Éléments de théorie analytique de l'information, modélisation et évaluation de performances

Philippe Jacquet

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Éléments de théorie analytique de l'information, Modélisation et évaluation de performances

Philippe Jacquet

No 3505

THÈME 1

ISSN 0249-6399

INRIA
ROCQUENCOURT

Éléments de théorie analytique de l'information, Modélisation et évaluation de performances

Philippe Jacquet

Thème 1 — Réseaux et systèmes
Projet Hipercom

Rapport de recherche n 3505 — — 86 pages

Résumé : Ce mémoire présente des résultats analytiques obtenus dans le domaine de la théorie de l'information, dont on fête le cinquantenaire cette année. Sont passés en revue, les calculs d'entropie, les algorithmes de compression, les structures de données, les réseaux de télécommunication et les réseaux sans fil. Dans chacun de ces domaines je présente des résultats qui utilisent des développements en analyse complexe.

Mots-clé : théorie de l'information, analyse complexe, entropie, compression, structures de données, protocoles de communication, réseaux sans fil

(Abstract: pto)

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : 01 39 63 55 11 - International : +33 1 39 63 55 11
Télécopie : (33) 01 39 63 53 30 - International : +33 1 39 63 53 30

Elements of analytical information theory, models and performance evaluation

Abstract: This memorandum presents analytical results which have been obtained in the domain of information theory, which celebrates its 50th anniversary this year. We review entropy calculations, the compression algorithms, the data structures, the telecommunication networks and the wireless networks. In any of these domains, I present results obtained via complex analysis.

Key-words: Information theory, complex analysis, entropy, compression, data structures,communication protocols, wireless networks.

Introduction générale

Un egoïste, c'est quelqu'un qui ne pense pas à moi.
(Anonyme)

Une thèse d'habilitation c'est plusieurs choses. D'abord c'est l'occasion d'un nouveau bizutage. Ou alors ça peut être aussi une sorte de franchissement en fanfare du point de Peter. De toute manière c'est l'occasion pour l'impétrant de faire l'exposé et le bilan de ses propres travaux. Afin de tenir la distance sur plus de cent pages à discourir, disserter, extrapoler sur ses réalisations propres, il est particulièrement important de rassembler au préalable une bonne dose d'égocentrisme. Sur ce plan là au moins, je pense avoir été à la hauteur du défi.

En premier lieu il m'a fallu trouver un fil conducteur qui permette aux lecteurs du présent mémoire, que je remercie d'avance, de ne pas se perdre dans le fouillis des publications. Ce fil conducteur m'a obligé à faire une sélection. En ne présentant qu'une facette de mon immodeste talent je ne veux pas renier les autres mais il valait mieux assurer la meilleure lisibilité possible.

Le fil conducteur que j'ai choisi est la théorie de l'information. Cela tombe particulièrement bien, parce 1998 est le cinquantenaire du papier fondateur de Shannon, "A mathematical Theory of Communication" [1] publié en juillet 1948. Depuis 1948 il est passé beaucoup d'eau sous les ponts, et il semble se confirmer que

1. l'informatique a tendance à se répandre de manière exponentielle;
2. que plus de 99% des usages des ordinateurs sont en fait limités à l'accès à et au traitement de l'information.

Les sciences de l'information ont donc gagné une importance de premier plan et sont même devenues des enjeux économiques, voire géopolitiques. Mais je n'ai pas voulu ici céder à un effet de mode, la théorie de l'information a ce quelque chose de fondamental et de maintenant bien ancré qui convient très bien au bilan que je veux faire ici. De plus c'est une science attirante qui s'est vraiment épanouie depuis le développement de l'informatique et que j'ai d'ailleurs personnellement un peu apprise sur le tas.

Je ne veux surtout pas renier mon berceau scientifique qui est l'algorithmique et l'analyse des algorithmes. Je suis heureux de souligner ici mon attachement à l'école initiée par Donald Knuth et brillamment développée par Philippe Flajolet. La théorie analytique de

l'information est une nouvelle petite branche que je voudrais apporter à cette école. Que nous voudrions apporter, car je ne voudrais pas oublier ici mon complice Wojciech Szpankowski avec lequel j'ai si souvent sévi dans les coulisses de la théorie de l'information (et dans celles du Crazy Horse accessoirement) et à qui on doit le terme de théorie analytique de l'information.

Bien sûr, je ne prétend pas avoir approché, même de très loin, l'intégralité du domaine de la théorie de l'information. Par exemple je n'ai jamais fourré mon nez dans la théorie de la cryptographie. Mon ami François Morain en a fait une perspective éblouissante et qui restera dans les mémoires (même si cela blesse un peu un peu mon égocentrisme de circonstance). Le domaine des codes correcteurs m'est aussi très arride malgré les efforts méritoires de Paul Camion ou Nicolas Sendrier pour m'en expliquer les subtilités. D'un autre coté j'ai beaucoup investi dans le domaine des télécommunications qui n'est qu'une petite partie de la théorie de l'information. C'est donc du petit bout de ma lorgnette que je vais essayer d'illustrer au mieux mon propos.

C'est à regret que j'ai dû écarter de la présente perspective mes différentes études sur la théorie des files d'attentes, les architectures de réseau et sur les algorithmes. Enfin il ne faut pas fatiguer le lecteur en faisant trop dispersé. Cela ne veut pas dire que je place au second plan ces domaines très interessants et très prometteurs.

Le premier chapitre de ce mémoire est consacrée à une introduction à la théorie de l'information. J'ai essayé d'éviter la tentation de faire dans le cosmique. J'ai néanmoins tenté de manière bien maladroite de replacer les sciences de l'information en perspectives avec les sciences fondamentales mathématisables. Un petit rappel original sur les fables de la Fontaine m'a permis de donner une touche un peu plus ludique à ce chapitre introductif.

Le second chapitre est consacré à la définition de l'entropie des sources d'information. Cette grandeur fondamentale qui mesure la quantité d'information est un peu l'équivalent de l'énergie en physique. Son calcul précis est crucial pour prédire les performances des systèmes de traitement de l'information. À titre d'illustration je présente deux papiers avec Wojciech Szpankowski, un sur la méthodologie de calcul liée à la poissonisation et un autre sur le calcul des entropies des distributions binomiales.

Le troisième chapitre est consacré à la compression d'information qui est un des domaines emblématiques de la théorie de l'information. La minimisation de la redondance par rapport à l'entropie est le premier défi des algorithmes de compression. Je présente deux papiers sur les algorithmes de Ziv et Lempel.

Le quatrième chapitre est consacré aux structures de données. Les structures de données sont historiquement les premiers espaces communs entre la théorie de l'information et l'informatique et sont naturellement en première ligne en ce qui concerne les techniques de stockage et d'accès à l'information. En illustration je présente deux papiers sur les *tries*, un avec Mireille Régnier et l'autre avec Wojciech Szpankowski. Je présente aussi la partie sur les arbres digitaux du papier avec Wojciech Spankowski déjà mentionné à l'occasion de l'analyse d'un algorithme de compression de Lempel et Ziv dans le chapitre précédent.

Le cinquième chapitre est consacré au transfert d'informations sur les réseaux et les modélisations de trafic de télécommunication. Ce chapitre est un peu plus grand que les autres

car il entre dans un domaine que j'ai exploré plus en profondeur. J'y montre comment de la loi de Shannon sur les capacités optimales des réseaux on peut dériver les capacités réelles et les capacités pratiques d'un système de télécommunication. La capacité pratique d'un réseau dépend du protocole de communication et du modèle de trafic. J'illustre ce chapitre avec des papiers sur les protocoles pour modems-câbles exposés au comité de normalisation IEEE 802.14 et avec un papier sur la modélisation des trafics Internet avec des sources *on/off*.

Le sixième et dernier chapitre est consacré aux réseaux de données sans fil. Ces réseaux ont connu un nouvel éclairage depuis l'émergence du réseau mobile GSM. L'étude des performances des réseaux sans fil est une branche spécifique des télécommunications qui se montre actuellement très active. Du point de vue de la modélisation, la capacité pratique d'un réseau sans fil dépend toujours du protocole et du trafic, mais s'ajoutent aussi les conditions de propagation des ondes. En fait on restreint le concept de capacité pratique au concept de capacité unitaire qui prend en compte la portée pratique des émission et la densité de trafic par unité de volume. Ce chapitre fait essentiellement échos aux études menées au sujet de la normalisation du réseau Hiperlan. J'y présente aussi une étude inédite sur les réseaux sans fil pour le contrôle aérien. Cette étude avait été menée dans le cadre d'une mission d'expertise auprès de l'Institut Européen de Standardisation des Télécommunications (ETSI).

Si vous avez tenu le coup jusque là, bravo! vous pouvez sans problème vous ligoter la suite. Bon courage!

Bibliographie

[1] C. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol 27, pp. 379-423, 623-656, Juillet-Octobre 1948.

Chapitre 1

Généralités

Au début était le Verbe, l'adjectif qualificatif épithète est venu après

1.1 Introduction

Tout a commencé le jour où l'homme a éprouvé le besoin de communiquer avec ses semblables. Quand l'homme a inventé le langage, l'Information était née; il a su se servir du feu, l'Énergie lui était révélée; et quand il a enfin su compter, l'Économie allait le faire prospérer.

D'une perspective mathématique les choses ne se sont pas passées dans le même ordre. On peut même dire que la science de l'Information est la dernière venue des sciences qui auront chatouillé la *gens mathematica*. En 1948, Claude Shannon a publié son article fameux "A Mathematical Theory of Communication," qui établissait la classification fondamentales des sciences de l'information. Au paravant la théorie de la communication était une branche du traitement du signal qui avait pris son essor depuis le siècle dernier [2]:

- Télégraphe (Morse 1830);
- Téléphone (Bell 1876);
- Télégraphe sans fil (Marconi 1887);
- Radio AM (1900);
- Télévision (1925);
- Radio FM (Armstrong 1936);
- Étalement de Spectre (1940).

Le papier de Shannon donnait en quatre parties les éléments fondamentaux de la nouvelle science.

La partie I présentait les cadres du domaine de la théorie mathématique des communications aux travers d'exemples contemporains.

La partie II établit les bases théoriques de la compression de données en quantifiant la quantité d'information contenue dans une source donnée et en présentant les algorithmes susceptibles d'assurer une compression optimale.

La partie III vise le problème de la capacité d'un canal de communication bruité pour transmettre sans erreur de l'information.

La partie IV s'interesse à la compression avec perte et au problème de la distorsion avec la source originale.

A l'exception de ceux de la partie IV, le présent mémoire va tenter de passer en revue chacun de ces points, et en rajouter d'autres en provenance de l'actualité scientifique.

Mais avant d'entrer dans les détails scientifique, on peut s'interroger comment une telle explosion des technologies de l'information a pu avoir lieu en si peu de temps, alors que le fondateur de la théorie est toujours notre contemporain. Pourquoi aussi la Théorie de l'Information est-elle restée un peu marginale, alors que le succès technologique est maintenant universel? Peut-être parce que l'Information reste un concept qui n'est pas aussi palpable que la matière et l'énergie. Mais est-ce suffisant?

Néanmoins si on accepte d'être cosmique encore juste une petite minute, voici comment on pourrait ordonner les sciences "exactes" dans l'ordre inverse de leur matérialité:

1. Les sciences de l'énergie et de la matière, inerte ou vivante;
2. les sciences de l'information et de la communication;
3. les sciences de l'économie et des moyens financiers.

En effet on évalue la totalité de la connaissance écrite humaine à quelques dizaines de milliards d'octets. Grace aux moyens informatiques modernes, on peut déplacer cette quantité d'information sur une fibre optique en moins d'une minute et en ne dépensant que quelques Joules. Néanmoins l'information est beaucoup plus matérielle que la monnaie. Le dollar a été la dernière monnaie en 1975 à abandonner sa convertibilité en or: le chryshendonisme[1] avait vécu. Désormais largement plus de 99% de la monnaie mondiale n'a pas plus d'existence tangible que quelques lignes dans les mémoires des ordinateurs dans les banques centrales. La preuve que l'argent est encore moins matériel que l'information: une ligne de crédit de 1 507 637 FF sur un relevé de compte peut être accidentuellement convertie en un déficit catastrophique si l'on place un signe "moins" devant et quelques zéros derrière, une modification négligeable au regard des lois de l'information.

Par un raccourci saisissant, la physique a sa célèbre formule $e = mc^2$ qui permet d'identifier l'énergie et la masse d'un élément en recourant à la géométrie de l'univers. Le $e = mc^2$ de l'information existe: c'est l'identification de l'entropie $h = -\sum_i p_i \log_2 p_i$ d'un support

1. *Chryshendonisme*: n.m. tendance qui prévalait chez les anciens de confondre moyens financiers et abondance de métaux précieux. *Synonyme*: picsoutite aiguë

avec le maximum d'information, exprimée en bit, qu'il peut contenir. Bien qu'elle utilise la même formulation, l'entropie de la théorie de l'information n'a pas la même finalité que l'entropie de la physique statistique, qui sert à mesurer le *désordre*.

En effet la seconde loi de la thermodynamique énonce que l'entropie physique de l'univers est vouée à augmenter. C'est une loi tellement inexorable qu'elle est souvent considérée comme la base de la détermination de la flèche de l'écoulement du temps, et accessoirement comme excuse classique pour les personnes désordonnées. Mais en information il en va tout autrement. Profitant du caractère immatériel de l'information, l'entropie de l'information est maléable à volonté et dans tous les sens sans craindre de représailles des physiciens. Ainsi les fragments d'information peuvent être au choix:

- *comprimés* si on veut épargner du support;
- *étirés* par l'adjonctions de codes correcteurs d'erreurs si on veut les protéger d'erreurs de transmission;
- *distordus* par du cryptage si on veut les garder confidentiels.

1.2 Éléments de base

Ici se place l'avertissement d'usage qui consiste à annoncer que l'on ne s'interessera à l'information que comme suite de caractères $x_1 x_2 \dots x_n \dots$ issus d'un alphabet fini sans référence à son signifiant. La quantité x_i dénote le $i$ème caractère du fragment d'information considéré. Sa valeur est un symbole dans l'alphabet. Le signifiant peut être indifféremment soit un traité d'astronomie, un livre de recettes de cuisines, le catalogue de la Redoute, ou un enregistrement 3D de la revue du Crazy Horse, cela n'aura aucune importance et ne sera pas pris en compte dans le présent mémoire. En d'autres termes la théorie de l'information s'arrête là où elle devient intéressante. Ces premiers principes admettent quelques exceptions notables:

- la suite discrète x_i est remplacée par une fonction réelle $f(t)$ d'une variable réelle t comme la fonction de balayage en télévision analogique;
- la suite x_i est remplacée par une suite bidimensionnelle $[x_{ij}]$, voire multi-dimensionnelle, comme celle du *bitmap* d'une image numérique.

Dans le présent mémoire, ces exceptions ne seront pas abordées dans la mesure où nous nous restreindrons à l'information numérique et que les suites bidimensionnelles pourront le cas échéant être décrites par des suites linéaires.

1.2.1 Modélisation des sources d'information

Le modèle le plus simple est le modèle dit de *Bernoulli* où les valeurs des caractères d'un texte, ou fragment d'information, sont choisies aléatoirement dans un alphabet fini et ce

indépendamment les unes des autres. Par exemple on peut choisir uniformément parmi les 26 lettres de l'alphabet plus le symbole correspondant au "blanc" pour écrire un texte.

ABCDEFGHIJKLMNOPQRSTUVWXYZ_

Un exemple de tirage uniforme est:

DRSFKFXUNQEJGRIBAGA FADDQGEVIFIXPROJOISGL

Dans l'exemple précédent on ne tient pas compte du caractère non uniforme de la répartition des lettres dans la langue française sans laquelle la lettre "Q" ne vaudrait pas huit fois la lettre "A" au *scrabble*. Dans le premier cas on a un modèle de Bernoulli uniforme où chaque lettre revient avec la même probabilité $\frac{1}{27}$. Si on veut être plus réaliste il faut tirer les lettres selon une statistique non-uniforme qui s'exprime par un vecteur de probabilité arbitraire $(p_1, \ldots, p_V)$ fixé *a priori* où p_i est la probabilité d'occurrence dans le texte aléatoire du $i$ème symbole d'un alphabet de taille V, $\sum_{i=1}^{i=V} p_i = 1$. C'est le modèle de Bernoulli non-uniforme. Quand, pour tout i, $p_i = \frac{1}{V}$ on revient au modèle de Bernoulli uniforme déjà décrit avant.

Par exemple, en nous servant de la statistique des lettres dans:

ELEMENTS DE THEORIE ANALYTIQUE DE L'INFORMATION

on effectue le tirage de Bernoulli non uniforme:

MENHEENO S ETETSYI STNLQ QYIQTAETLETTO T

Un pas supplémentaire important est franchi vers davantage de réalisme quand on utilise des modèles dits de *Markov*. Dans le modèle de Markov le choix d'un caractère dépend du caractère précédent. Par exemple dans la langue française la lettre "Q" a beaucoup plus de chance d'être suivie par la lettre "U" que par la lettre "X". Pour son initialisation un modèle de Markov nécessite l'identification du premier caractère du texte à créer plus une matrice carrée de transition $V \times V$: $[p_{ij}]$ où p_{ij} est la probabilité d'avoir le caractère égal au symbole j quand le le caractère précédent a pour valeur i. Pour tout i: $\sum_{j=1}^{j=V} p_{ij} = 1$. Si pour i variant, les vecteurs $(p_{i1}, \ldots, p_{iV})$ sont tous identiques on en revient au modèle de Bernoulli.

En utilisant la statistique des transitions dans le texte échantillon:

THEORIE ANALYTIQUE DE L'INFORMATION

on effectue le tirage de Markov, en partant de la même lettre "T" (mise en boite):

TIQUE DE LYTHE L'IORMAL'IE LYTHE DE

Le modèle de Markov améliore le modèle de Bernoulli grâce à son principe de mémoire. Une portion d'un texte y dépend de manière statistique des portions qui le précèdent. Dans le modèle de Bernoulli les portions de texte sont indépendantes les unes des autres: les modèles de Bernoulli sont sans mémoire. En fait, le modèle de base de Markov décrit ci-dessus ne

nécessite qu'une mémoire de taille un: il suffit de connaître le dernier caractère du texte déjà créé pour prédire la statistique du texte qui suit.

On peut généraliser les modèles de Markov à des mémoires de profondeur supérieure: deux, trois ou plus, M. Dans ce cas le modèle s'exprime au travers d'un mot initial de taille M et d'une matrice de transition $V^M \times V$: $[p_{aj}]$ où p_{aj} désigne la probabilité qu'un sous mot a de taille M soit suivi du symbole j.

Par exemple lorsque part des statistiques des transitions à mémoire 2 du texte échantillon on obtient, par un tirage de Markov à mémoire 2:

TH EORMATIQUE DE DE ANALYTIQUE ANALY

La même chose en mémoire 3 redonne le texte original:

THE ORIE ANALYTIQUE DE L'INFORMATION

Une nouvelle étape dans la généralisation des modèles est le modèle *à renouvellement*, dans lequel le processus passe par des phases ponctuelles de renouvellement ou le futur ne dépend plus du passé. La généralisation ultime est le modèle ergodique qui est une sorte de modèle minimal où il est seulement exigé que tous les états atteignables par la source d'information le sont une infinité de fois sur un texte infini [5, 6, 7]. Dans ce qui suit nous nous restreindrons volontairement aux modèles de Bernoulli et de Markov.

1.3 Un jeu de Shannon

Afin d'illustrer notre propos nous nous sommes livrés à la petite expérience suivante inspiré d'un jeu de Shannon. Nous avons pris un texte fameux de la littérature francaise et nous avons établi la statistique de la valeur de ses caractères (en omettant accentuation et ponctuation). À partir de cette statistique nous avons produit un texte de même longueur par le modèle de Bernoulli.

Ensuite sur le texte initial nous avons établi la statistique des transitions de Markov selon une mémoire de taille un, et créer un nouveau texte de même longueur selon le modèle de Markov et cette nouvelle statistique.

Ensuite nous avons fait de même pour des statistiques à mémoire M arbitraire et à chaque fois nous avons créé un texte de même longueur selon le modèle de Markov correspondant à la mémoire de taille M.

Dans ce qui suit nous avons rangé les textes aléatoires par ordre de mémoire croissante, le jeu consistant à reconnaître le texte d'origine le plus rapidement possible (sans tricher en regardant à partir de la fin).

Mémoire de profondeur 0

RAAS IELC' O AEU
AOS
TLD ARFPLRTEE FAEP
V SUEHELLESNEA'OATC STTQAA EAQEEUZEPTOUETALEFIEHFUSEZELPEJTOIPIRN-
GEET
HLTE OE' STSLENE
LTAE U

NAPEDSEBI SESESN OTO REMDOUUOFJANIRP
C HHE AGAINEALJE EPFPPIFSSAIU EIEMTSORI
EEUESTAOGD A NUELLQEEIT'T REQVEH SNENE TS ACOSSEU NAESDUAV FRI
S TEN MLPN LETEOUE
D N DTSMAAE C L ER TEYSAT TUNA SZFTPAZ OLMAENPARTLOTUARU VHDZA
UUENR CLI
T D C SSETVAD Q NLL
DAUTTNOIO FPVIERSLF NE UCRAT IEREEEOLSNIUE'IDITTPLEAIRSUTNOES PEAS
AT AAHVR
L SA E IVAVEDUEASMEFM
ETVSNT ELIUNTUUR EI SUSQV ASTSTTLUEN S JANIETNN TSSEE TT
SS OCI
ASJAEN

Mémoire de profondeur 1

lquseri fou'a fais ve lll font l'oinera let pan sout lletite sure
selqurisu prmoitrt voue chasta let cisa cise d vour
ce
s cha mous monere
chavubi c'a cheant delurt toritis vor vez gasi prt tese
crtan mouer plle
dez prona die van lllqus prchaninte
n suepa cha valetisa de
lanigrchema det faians vue foinde ler
que lepraue pai pa one voud
quvoubial
quteseua fai vourientendaniplepa pr
pami di stis de e
vor
qun'e bipe
vorinous pr disermoner velut
e
chave
qubisemaie
d'a al'e ainet vus derte
e jemiset pr tte da e pavu jousetinale t ousellet llanourvouie
crmpansoubite len la

Mémoire de profondeur 2

LA SUIT ETTERAISISSE
AVANT PRETIER SUI DANTEUS CHE ALLE OURMINE
JE FAMI DEPLA VER
JE DANT DE AIER
QUE
QUE
LAI D'A SE
LA FAINE
ELLA PAS ALE AYANSEAU DE FOINCIGALLE EMPRIANTER AYANT LA SA TOURVUE
PAS CIPAS PRIND LA PAIND L'EST ELLA CIGALE FUT ELLA FUT
QUE
LA CHANTE
VOURVUE GRAI SUBSIN MOI PRE DANTIT DEFAMINT
QU'A LUI N'EST MORCEAU
ELLE
C'EST CHANSEUSE
PAS A CHANSE
VOUSEAU DIT VENUELLE VEN PETE
AVA L'OURMI PRETENUI PRET J'EN PRIER
QUE FOUR SAISE CHANT POURVUELLE LA CIPALE CHANT ETEMPS UN SA TENUIS
DE AIER FOUSE VOUVELLE VENANT EMPS AU DANTE
QUE AL
LA SA LA FORT DEPLAIER

Mémoire de profondeur 3

LA PRET PAS PAS UN SUBSISTERETE
TOUT VENANT
LA SAISON MORCEAU
ELLE AYANT
LA SON MOINDRE DEPOURMI N'EST L'OUT VENUE
PAS UN SUIS NE VOISINE
CHEZ J'EN DANSEZ MAIN POURMI SAIS NE VOISIEZ VOISINE
LA FORT AISINE
CHEZ LA CIGALE A TOUT VERMI N'EST L'OUT L'ETER
QUELQUE GRAINTERETEUSE
VOUS DEPOUR SUIS FOI D'ANIMAL
INTE
TOUT LA CIGALE AYANTE
TOUT LA CIGALE ALLA SA VOUS PRIANT
JE CHANT CHAUD
DIT ET JOURMI N'EST L'ETEUSE
NUIT ET PAIERAI LUI D'ANIMAL
INTEUSE
EH BIEN SUBSISTERETE
SE FAMINE
LA SA VOISIEZ VOISINE
CHE OU DEFAUT
QUELQUE GRAI LUI PRIANTIEZ J'EN SEUL PETIT ET ET PRETE
SE

Mémoire de profondeur 4

LA CIGALE AYANT L'OUT FOI D'ANIMAL
INTENANT
LA CRIER FAMINE
LA CIGALE A CETTE EMPRUNTEUSE
NUIT ELLE
JE VOUS PAIERAI LUI DIT ET PRIANT DE VERMISSEAU
ELLE
JE VOUS DEPLAISE
VOUS CHANTAIS NE VOUS DEPOURVUE
QUAND LA BISE FUT VENANT
JE CHAUD
DIT ELLE
AVANT DEPOURVUE
QUAND LA FOURMI SA VOISINE
LA PRIANT L'ETE
SE TROUVA FOURMI N'EST LA SAISON MOINDRE DEFAUT
QUE FAISIEZ VOUS AU TEMPS CHAUD
DIT ET PRIANT CHANTE
TOUT FOI D'ANIMAL
INTENANT
LA CRIER FAMINE
CHEZ LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE AYANT DEPLAISE
VOUS DEPLAISE
VOUS PAIERAI LU

Mémoire de profondeur 5

LA CIGALE AYANT CHANTE
TOUT L'ETE
SE TROUVA FORT AISE
VOUS CHAUD
DIT ELLE ALLA CRIER FAMINE
CHEZ LA FOURMI N'EST LA SON MOINDRE DEFAUT
QUE FAISIEZ VOUS AU TEMPS CHANTAIS NE VOUS DEPLAISE
VOUS AU TEMPS CHANTIEZ J'EN SUIS FORT DEPOURVUE
QUAND LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE A CETTE EMPRUNTEUSE
NUIT ET PRINCIPAL
LA FOURMI N'EST PAS PRETEUSE
C'EST PAS PRETEUSE
C'EST LA SAISON NOUVELLE
JE VOUS CHANTE
TOUT L'ETE
SE TROUVA FORT AISE
VOUS AU TEMPS CHANTIEZ J'EN SUIS FORT AISE
EH BIEN DANSEZ MAINTENANT
LA CIGALE AYANT CHANTIEZ J'EN

Mémoire de profondeur 6

LA CIGALE AYANT CHANTIEZ J'EN SUIS FORT AISE
EH BIEN DANSEZ MAINTENANT
LA CIGALE AYANT CHANTIEZ J'EN SUIS FORT DEPOURVUE
QUAND LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE ALLA CRIER FAMINE
CHEZ LA FOURMI SA VOISINE
LA PRIANT DE LUI PRETER
QUELQUE GRAIN POUR SUBSISTER
JUSQU'A LA SAISON NOUVELLE
JE VOUS AU TEMPS CHAUD
DIT ELLE
AVANT L'OUT FOI D'ANIMAL
INTERET ET PRINCIPAL
LA FOURMI SA VOISINE
LA PRIANT DE LUI PRETER
QUELQUE GRAIN POUR SUBSISTER
JUSQU'A LA SAISON NOUVELLE
JE VOUS DEPLAISE
VOUS CHANTE
TOUT L'ETE
SE TROUVA FORT AISE
EH BIEN

Mémoire de profondeur 7

LA CIGALE AYANT CHANTE
TOUT L'ETE
SE TROUVA FORT DEPOURVUE
QUAND LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE ALLA CRIER FAMINE
CHEZ LA FOURMI SA VOISINE
LA PRIANT DE LUI PRETER
QUELQUE GRAIN POUR SUBSISTER
JUSQU'A LA SAISON NOUVELLE
JE VOUS DEPLAISE
VOUS CHANTIEZ J'EN SUIS FORT AISE
EH BIEN DANSEZ MAINTENANT
LA CIGALE AYANT CHANTE
TOUT L'ETE
SE TROUVA FORT DEPOURVUE
QUAND LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE ALLA CRIER FAMINE
CHEZ LA FOURMI SA VOISINE
LA PRIANT DE LUI PRETER
QUELQUE GRAIN POUR SUB

Mémoire de profondeur 8

LA CIGALE AYANT CHANTE
TOUT L'ETE
SE TROUVA FORT DEPOURVUE
QUAND LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE ALLA CRIER FAMINE
CHEZ LA FOURMI N'EST PAS PRETEUSE
C'EST LA SON MOINDRE DEFAUT
QUE FAISIEZ VOUS AU TEMPS CHAUD
DIT ELLE
AVANT L'OUT FOI D'ANIMAL
INTERET ET PRINCIPAL
LA FOURMI N'EST PAS PRETEUSE
C'EST LA SON MOINDRE DEFAUT
QUE FAISIEZ VOUS AU TEMPS CHAUD
DIT ELLE
AVANT L'OUT FOI D'ANIMAL
INTERET ET PRINCIPAL
LA FOURMI N'EST PAS PRETEUSE
C'EST LA SON MOINDRE DEFAUT
QUE FAISIEZ VOUS AU TEMPS CHAUD
DIT ELLE
AVANT L'OUT FOI D'ANIMA

Mémoire de profondeur 9

LA CIGALE AYANT CHANTE
TOUT L'ETE
SE TROUVA FORT DEPOURVUE
QUAND LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE ALLA CRIER FAMINE
CHEZ LA FOURMI N'EST PAS PRETEUSE
C'EST LA SON MOINDRE DEFAUT
QUE FAISIEZ VOUS AU TEMPS CHAUD
DIT ELLE A CETTE EMPRUNTEUSE
NUIT ET JOUR A TOUT VENANT
JE CHANTAIS NE VOUS DEPLAISE
VOUS CHANTIEZ J'EN SUIS FORT AISE
EH BIEN DANSEZ MAINTENANT
LA CIGALE AYANT CHANTE
TOUT L'ETE
SE TROUVA FORT DEPOURVUE
QUAND LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE ALLA CRIER FAMINE
CHEZ LA FOURMI S

Mémoire de profondeur 10

LA CIGALE AYANT CHANTE
TOUT L'ETE
SE TROUVA FORT DEPOURVUE
QUAND LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE ALLA CRIER FAMINE
CHEZ LA FOURMI SA VOISINE
LA PRIANT DE LUI PRETER
QUELQUE GRAIN POUR SUBSISTER
JUSQU'A LA SAISON NOUVELLE
JE VOUS PAIERAI LUI DIT ELLE
AVANT L'OUT FOI D'ANIMAL
INTERET ET PRINCIPAL
LA FOURMI N'EST PAS PRETEUSE
C'EST LA SON MOINDRE DEFAUT
QUE FAISIEZ VOUS AU TEMPS CHAUD
DIT ELLE A CETTE EMPRUNTEUSE
NUIT ET JOUR A TOUT VENANT
JE CHANTAIS NE VOUS DEPLAISE
VOUS CHANTIEZ J'EN SUIS FORT AISE
EH BIEN DANSEZ MAINTENANT

1.4 Exercice commenté: la profondeur d'un texte

Le jeu précédent met en lumière la notion de profondeur maximale d'un texte. C'est à dire la profondeur maximale de la mémoire du processus de Markov au delà de laquelle le texte aléatoire reste identique au texte échantillon original. En l'occurrence il a suffit d'une mémoire de profondeur 10 pour retrouver le texte intégral de la *La Cigale et la Fourmi* de Jean de la Fontaine.

En fait le texte initial ne contient pas de segments de 10 caractères consécutifs qui soit dupliqué. En d'autres termes la connaissance de 10 caractères consécutifs de la Cigale et la Fourmi suffit pour déterminer la valeur du onzième caractère. Par exemple, la séquence "LA CIGALE A" ne peut etre suivie que par la lettre "Y" puis la séquence "A CIGALE AY" ne peut suivie que par la lettre "A", et ainsi de suite jusqu'à donner "LA CIGALE AYANT CHANTÉ...".

La profondeur maximale est évidemment fonction du texte. Un texte original constitué d'une série de caractères tous différents, comme par exemple:

ABCDEFGHIJKLMNOPQRSTUVWXYZ

donne lieu à une profondeur 1. Si par contre le texte original est constitué d'une série de caractères tous identiques, comme par exemple:

AAAAAAAAAAAAAAAAAAAAAAAAA

alors la profondeur maximale est égale à la longueur du texte.

Pour un texte original aléatoire de longueur n, nous appelons H_n sa profondeur (maximale). L'objectif du présent exercice est de calculer une borne supérieure de la distribution de H_n lorsque le texte est soumis au modèle de Bernoulli sans mémoire. Un texte créé par un processus de Markov de profondeur zero ne donne pas lieu à un texte de profondeur maximale zéro et en général $H_n \gg 0$. Nous allons montrer qu'en probabilité $H_n = O(\log n)$. Ce résultat a été montré dans [5] et [18].

La quantité H_n est égale à la longueur plus une unité du plus grand fragment de texte qui soit dupliqué dans une autre partie du texte (les superpositions sont permises). Recopions sur deux segments différents les deux suffixes du texte original commençant respectivement au i-ème et j-ème caractères du texte. On appelle $C(i,j)$ la longueur du plus grand préfixe commun aux deux segments recopiés. Il est clair que

$$H_n = \max_{i,j}\{C(i,j)\} + 1 \ . \tag{1.1}$$

De l'identité précédente, on tire l'inégalité, valable pour tout entier k:

$$\Pr\{H_n \geq k\} \leq \sum_{i,j \leq n} \Pr\{C(i,j) \geq k-1\} \ . \tag{1.2}$$

Par symétrie de translation on a $\Pr\{C(1,1+\ell) \geq k\} = \Pr\{C(i,\ell+i) \geq k\}$ pour tout $i \leq n-\ell$. Pour éviter de traîner de bout en bout le fardaud des effets de bord avec l'extrémité droite du texte nous supposerons que le texte original est simplement infini à droite tout en ne considérant que ses n premiers suffixes. Remarquons que cette simplification respecte l'inégalité (1.2). En reprenant cette dernière on obtient la nouvelle inégalité:

$$\Pr\{H_n \geq k\} \leq \sum_{\ell=1}^{\ell=n-1} (n+1-\ell)\Pr\{C(1,1+\ell) \geq k\} \ . \tag{1.3}$$

La détermination des quantités $\Pr\{C(1,1+\ell) \geq k\}$ nécessite l'étude de deux cas de figure: $k < \ell$ et $k \geq \ell$.

Le cas $k < \ell$ est le plus simple dans la mesure où les segments de texte à comparer ne se chevauchent pas. La quantité $\Pr\{C(1,\ell+1) \geq k\}$ est égale à la probabilité que deux segments disjoints de longueur k soient identiques dans le modèle de Bernoulli. On a donc $\Pr\{C(1,\ell+1) \geq k\} = (|p|_2)^k$, avec la notation $|p|_m = \sum_{i=1}^{i=V} p_i^m$.

Le cas $k \geq \ell$ est un peu plus délicat à traiter puisque les segments à comparer se chevauchent sur une longueur $k-\ell$.

Pour simplifier regardons tout d'abord le cas où ℓ divise k: $k = q\ell$ pour q entier. Appelons $X(1,\ell)$ le fragment de texte situé entre la position 1 et ℓ du texte initial. L'hypothèse

$C(1,1+\ell) \geq k$, est eéquivalente au fait que le texte original commence par le mot $X(1,\ell)$ répété $q+1$ fois. Cette dernière propriété implique que toutes les lettres situées dans cet intervalle et dans des positions de rangs identiques *modulo* ℓ sont identiques. Pour chaque valeur du *modulo* ℓ il y a exactement $q+1$ positions de rang identique *modulo* ℓ, donc l'égalité entre ces lettres se fait avec une probabilité cumulée de $|p|_{q+1}$. Comme il y a ℓ positions différentes *modulo* ℓ on obtient donc $\Pr\{C(1,1+\ell) \geq q\ell\} = (|p|_{q+1})^{\ell}$.

Dans le cas où ℓ ne divise pas k on a une variante du raisonnement précédent où interviennent le quotient q de la division k/ℓ et son reste r. On a toujours la propriété que dans l'intervalle $(1, k+\ell+1)$ toutes les lettres sur des positions de rangs identiques *modulo* ℓ sont identiques. Mais maintenant le nombre de positions de rang identique *modulo* ℓ est soit $q+1$ ou $q+2$, dépendant de la valeur du *modulo* dans l'intervalle $(1,\ell)$. En fait il y a $q+2$ positions pour les valeurs de *modulo* inférieurs ou égaux à r et $q+1$ positions pour les $\ell - r$ autres valeurs du *modulo*. On obtient donc

$$\Pr\{C(1,1+\ell) \geq k\} = (|p|_{q+2})^{r}(|p|_{q+1})^{\ell-r} \ . \tag{1.4}$$

Dans le modèle de Bernoulli uniforme on a $|p|_m = V^{1-m}$, où V est la taille de l'alphabet. Donc $\Pr\{C(i,j) \geq k\} = V^{-k}$ quels que soient (i,j). Donc avec le modèle uniforme:

$$\Pr\{H_n \geq k\} \leq \binom{n}{2} V^{-k} \ . \tag{1.5}$$

Il apparaît donc:

1. la probabilité marginale, $\Pr\{H_n \geq \frac{\log(n^2/2)}{\log V} + x\} \leq V^{-x}$;
2. l'espérance mathématique $\mathrm{E}H_n \leq \frac{2\log n - \log 2}{\log V} + \frac{V}{V-1}$.

En ce qui concerne le modèle Bernoulli général on introduit les notations suivantes: $\max(p)$ et $\max_2(p)$ désignent respectivement la plus grande valeur et la seconde plus grande valeur atteinte par les quantités p_i quand i décrit l'intervalle $(1,V)$. On utilise l'estimation suivante:

$$|p|_m = \max(p)^m + O(V\max_2(p)^m) \ .$$

Donc $\Pr\{C(1,1+\ell) \geq k\} \leq \max(p)^{k+1} + O(V^2\max_2(p)^{k+1})$. On en tire donc l'inégalité:

$$\Pr\{H_n \geq k\} \leq \binom{n}{2}(\max(p)^{k+1} + O(V^2\max_2(p)^{k+1})) \ . \tag{1.6}$$

De l'expression précédente on tire les propriétés asymptotiques suivantes:

1. la probabilité marginale, $\Pr\{H_n \geq \frac{\log(n^2/2)}{-\log\max(p)} - 1 + x\} \leq (1+o(1))\max(p)^x$;
2. l'espérance mathématique $\mathrm{E}H_n \leq \frac{2\log n - \log 2}{-\log\max(p)} + \frac{\max(p)}{1-\max(p)} + o(1)$.

Si la valeur $\max(p)$ est atteinte plusieurs fois dans le vecteur p_i, par exemple D fois, alors $|p|_m = D\max(p)^m + O(V\max_2(p)^m)$. Dans ce cas on retrouve un facteur correctif dans $\Pr\{H_n \geq \frac{\log(n^2/2)+\log D}{-\log\max(p)}\}$ et $\mathrm{E}H_n \leq \frac{2\log n - \log 2 - \log D}{-\log\max(p)} + \frac{\max(p)}{1-\max(p)} + o(1)$.

Bibliographie

[1] S. Verdu, "Fifty years of Shannon Theory," preprint, juin 1998.

[2] J. Ziv, "Coding theorems for individual sequences," |it IEEE Trans. on Inform. Theory, vol 24, pp. 405-412, 1978.

[3] T. Cover, J. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[4] A. Wyner, J. Ziv, "The sliding-window Lempel-Ziv algorithm is asymptotically optimal," Proc. IEEE, pp. 872-877, 1994.

[5] L. Devroye, B. Rais, W. Szpankowski, "A note on the height of suffix trees," *SIAM J. Computing*, 21, pp. 48-53, 1992.

[6] W. Szpankowski, "A generalized suffix tree and its (un)expected asymptotic behaviors," *SIAM J. Computing*, 22, pp. 1176-1198, 1993.

Chapitre 2

L'entropie de l'information

De la célèbre Tour de Babel qui défiait les cieux, il ne reste que le Babylomètre: toise pour mesurer le verbiage

2.1 Introduction

La quantification de l'information est une préoccupation qui remonte à très loin. Déjà à la Renaissance, l'analyse statistique des sources d'informations suscitait un grand intéret. À cette époque la diplomatie avait pris une telle importance que les casseurs de codes et de messages secrets en tout genre faisaient feux de tout bois pour arriver à leurs fins. Quelques unes des tables en usages dans ces officines sont parvenues jusqu'à notre époque. Ces tables contiennent les fréquences des symboles dans les langues usuelles et certaines remontent à 1380 et 1658 [1].

C'est à Shannon que revient le mérite d'avoir développé la notion d'entropie dans la théorie de l'information. Auparavant, certains chercheurs, comme Nyquist en 1924 [2] et Hartley en 1928 [3], avaient eu l'intuition d'utiliser le logarithme du nombre de sélections dans une source pour quantifier l'information

2.1.1 La mesure de l'incertitude

L'entropie d'une variable aléatoire sert à mesurer la quantité d'incertitude liée à cette variable aléatoire. Par exemple une variable binaire qui donne 0 et 1 avec la même probabilité 0.5 présente plus d'incertitude que la variable biaisée qui donnerait 0 avec probabilité 0.99 et 1 avec probabilité 0.01. On mesure l'entropie d'une variable aléatoire discrète X par l'expression

$$h(X) = -\sum_i p_i \log p_i \tag{2.1}$$

où i décrit l'ensemble des valeurs atteignables par X. Une formule courte mais abusive est $h(X) = -\mathrm{E}[\log p_i]$ où E[.] désigne l'expérance mathématique. La formule est abusive en ce sens que la distribution p_i de la variable X ne constitue pas une variable aléatoire à proprement dite. Dans l'exemple de la variable binaire on obtient $h(X) = 0.6931\ldots$ dans le cas uniforme qui domine la valeur de l'entropie $0.0560\ldots$ obtenue dans le cas biaisé.

Un autre exemple est donné par la comparaison des entropies de la distribution uniforme $p_i = \frac{1}{n+1}$ et de la distribution binomiale $p_i = \binom{n}{i} 2^{-i}$. Cette dernière étant plus "piquée" sur sa moyenne (voir figure 2.1), son entropie est plus faible. C'est ce que confirme les chiffres: pour $n = 20$, entropie uniforme $h = 3.044522$, entropie binomiale $h = 2.223423$.

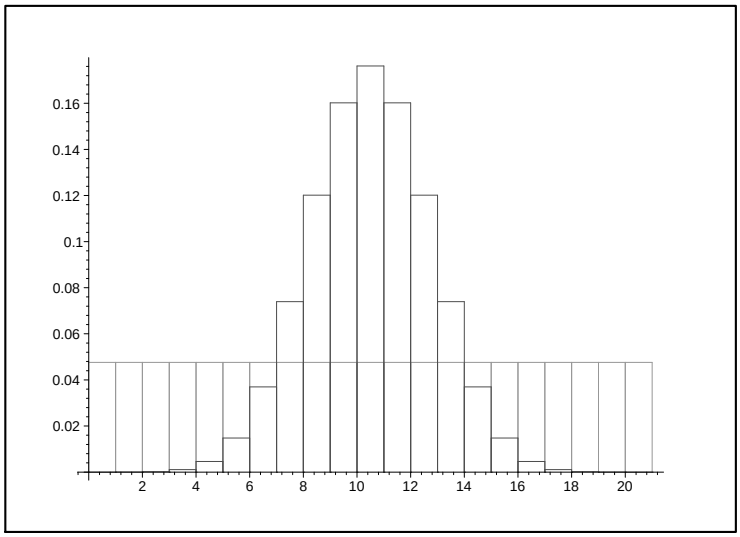

FIG. 2.1 – *Distribution uniforme et distribution binomiale*

Si $f(X)$ est une fonction de X, on a

$$h(f(X)) \leq h(X) \tag{2.2}$$

et on a égalité lorsque $f(.)$ est injective.

Pourquoi l'entropie mesure-t-elle la quantité d'information? Il existe de nombreuses interprêtations de l'entropie en théorie de l'information dont certaines sont liées à la seconde

loi de la thermodynamique. Dés le début Shannon exploita les analogies avec l'entropie de Boltzmann en physique statistique [4]. En particulier l'entropie est une fonction continue de la distribution de probability, elle croît lorsque le support de la variable aléatoire (nombre de facteurs p_i) augmente.

Plus important l'entropie présente d'intéressantes propriétés additives sur les distributions jointes. Prenons deux variables aléatoires X et Y, l'entropie de la variable aléatoire jointe est $h(X,Y) = -\sum_{ij} p_{ij} \log p_{ij}$. On définit l'entropie *conditionnelle* $h(Y|X)$ par

$$h(Y|X) = -\sum_i p_i h(Y|X=i) \tag{2.3}$$

où $(Y|X=i)$ désigne la variable aléatoire Y conditionnée par l'événement $X=i$. Par des calculs simples on arrive à l'identité appelée "règle de la chaîne":

$$h(X,Y) = h(X) + h(Y|X)\ . \tag{2.4}$$

Il découle de ce qui précède que l'entropie de deux variables indépendantes X et Y est égale à la somme des deux entropies séparées: $h(X,Y) = h(X) + h(Y)$.

2.1.2 L'entropie d'une source d'information

L'entropie d'une source d'information peut être interprétée comme une mesure de la quantité d'information qu'elle est susceptible de transporter. En effet plus la source sera "incertaine" plus elle sera capable de transporter de la variété. Par exemple une variable binaire uniforme pourra dire deux choses (1 ou 0) alors que la variable biaisée telle que décrite plus haut, ne dira guère autre chose que 0.

Prenons comme exemple un canal de communication, ou n'importe quel autre support d'information, susceptible de contenir tous les textes binaires de taille n de manière équiprobable. Il y a 2^n textes binaires de longueurs n. L'entropie de cette source d'information est l'entropie de la variable aléatoire X_n qui correspond à un texte binaire de Bernoulli avec distribution uniforme. En conséquence la capacité de ce support est $h(X_n) = -\sum_{\text{texte}} \Pr\{\text{texte}\} \log \Pr\{\text{texte}\}$. L'unité de mesure de l'entropie est dans certaines littératures désignée sous le nom barbare de *logons*. L'expression de l'entropie peut être divisé par $\log 2$ si l'on désire mettre en évidence la capacité *binaire* (en bit) de la source d'information. La capacité V-aire (obtenue avec un alphabet de taille V) s'obtient en divisant l'entropie par $\log V$.

Notons que la fonction $h(X_n)$ est croissante avec n si le texte X_n est le préfixe de taille n d'un texte aléatoire infini.

En utilisant la propriété d'additivité sur les variables indépendantes l'entropie d'une source sous modèle Bernoulli a pour expression:

$$h(X_n) = nh(x) = -n \sum_i p_i \log p_i \tag{2.5}$$

La quantité $h(x)$ désigne l'entropie d'un symbole x isolé.

Si tous les textes binaires sont équiprobables on obtient $h = n$ bits. Si au lieu d'un texte binaire on regarde un texte écrit dans un alphabet de taille V et que l'on considère toujours les textes équiprobables on obtient $h = n\frac{\log V}{\log 2}$.

Theorem 1 *Pour une taille d'alphabet donné V et une longueur de texte donnée, n, l'entropie atteint son maximum lorsque le texte est Bernoulli uniforme: en logons*

$$h(X_n) \leq n \log V \ . \tag{2.6}$$

La figure 2.2 montre la courbe de l'entropie d'une variable binaire en fonction de la première composante p du vecteur de distribution.

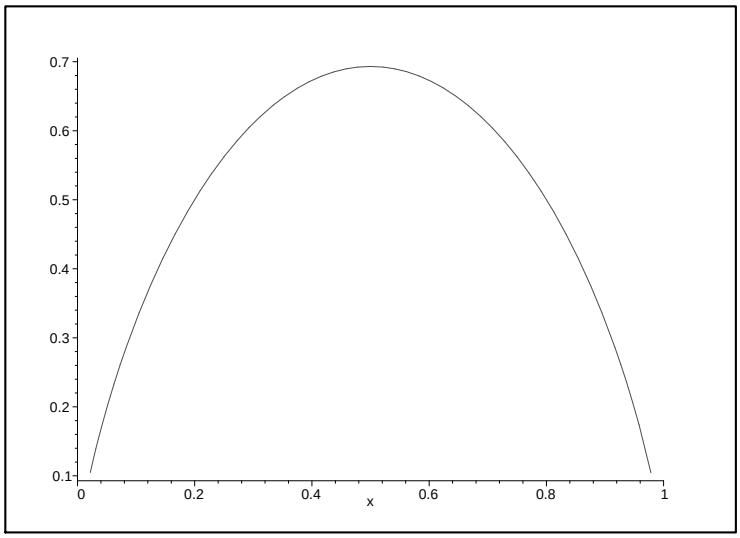

FIG. 2.2 – *courbe de l'entropie pour une variable binaire.*

L'entropie est additive vis à vis de deux familles de textes concaténés de manière indépendante: $h(\text{texte}_1 * \text{texte}_2) = h(\text{texte}_1) + h(\text{texte}_2)$.

La plupart des modèles de textes aléatoires sont *mélangeants* c'est à dire l'entropie ajoutée par le n-ème caractère tend à devenir indépendante de n. En d'autres termes il existe h tel

que

$$h = \lim_{n \to \infty} (h(X_{n+1}) - h(X_n)) \tag{2.7}$$

La quantité h s'appelle la *densité* d'entropie. Les modèles de Bernoulli et de Markov sont mélangeants. Un exemple de modèle non mélangeant est le suivant: le texte est issu d'un texte de Bernoulli entre les caractères duquel on intercale un symbole fixe. Dans ce cas $h(X_{2n+1}) = h(X_{2n})$ pour tout n.

Il existe des études poussées sur l'évaluation des langues usuelles, comme l'Anglais [9, 10, 11]. Par exemple l'entropie de la langue anglaise employée dans *Jefferson the Virginian* a été mesurée à 1.34 bits par lettre, soit 0.93 logons par lettre.

2.1.3 La capacité d'un canal bruité

Nous introduisons la notion de canal. Un canal transforme un texte d'entrée X en un texte de sortie Y qui est une fonction aléatoire de X. La capacité d'un canal soumis à un texte source X et délivrant un texte de sortie Y est notée $I(Y;X)$ et a pour expression:

$$I(Y;X) = h(Y) - h(Y|X) \tag{2.8}$$

La capacité d'un canal est donc la mesure de la quantité d'information qui le traverse, elle est obtenu en soustrayant à l'entropie de la sortie $h(Y)$ l'entropie conditionnelle $h(Y|X)$ provoquée par le bruit du canal. Pour des raisons de convexité ce résultat est toujours supérieur ou égal à zéro.

Capacité d'un canal binaire Soit un canal binaire dont la sortie prend des valeurs -1 ou $+1$. Nous supposons que le canal est soumis à deux processus aléatoires U_0 et U_1 à la discrétion d'un opérateur désirant transmettre de l'information. Dans le premier processus aléatoire les -1 et les $+1$ sont produits de manière indépendante selon les probabilités respectives (p_0, q_0). Dans le second processus les -1 et les $+1$ sont produits de la même manière mais avec les probabilités respectives (p_1, q_1). Nous allons montrer que la capacité du canal par bit transmis lorqu'on transmet de l'information en alternant les processus U_1 et U_2 est le maximum de la fonction

$$g(xp_0 + (1-x)p_1) - xg(p_0) - (1-x)g(p_1) \tag{2.9}$$

pour x variant dans l'intervalle $[0,1]$, où $g(y) = -y \log y - (1-y)\log(1-y)$.

Supposons que les bit 0 soient modulés avec U_0 et les bit 1 avec U_1. On a $h(Y) = g(xp_0 + (1-x)p_1)$ où x désigne la proportion des bits transmis sous processus U_0 et $1-x$ la proportion des bits transmis sous U_1. L'entropie du processus U_0 est $g(p_0)$, et celle de U_1 est $g(p_1)$.

L'entropie du texte résultant après modulation par U_1 et U_2 est donc $g(xp_0 + (1-x)p_1)$. En utilisant la formule de la capacité on obtient donc le résultat recherché.

Donc la capacité du canal est obtenue en maximisant la quantité $g(xp_0 + (1-x)p_1) - xg(p_0) - (1-x)g(p_1)$. Si $p_1 = 0$ et $p_2 = 1$ la capacité par bit est de 1, obtenue pour $x = 1/2$.

Si $p_1 = 0.1$ et $p_2 = 0.9$, équivalent à un "taux d'erreurs" symétrique de 10%, on obtient un effondrement de près de moitié de la capacité, précisément $1 + \log_2 0.1 + \log_2 0.9 \approx 0.531$

2.2 Exercice commenté

Entropie d'une source markovienne A. Khinchin [6] a analysé en détail les propriétés de l'entropies des chaînes de Markov de mémoire arbitraire. Pour la mémoire d'ordre 1, on montre que l'entropie d'un texte de longueur n obéissant au modèle de Markov est asymptotiquement équivalente à $-n \sum_{ij} \Pi_i p_{ij} \log p_{ij}$ où p_{ij} sont les coefficients de la matrice de transition probabiliste et Π_i sont les composants du vecteur stationnaire de cette matrice. La preuve applique la règle de la chaîne. Si X_n désigne le texte de longueur n on a la récurence:

$$h(X_{n+1}) = h(X_n) + h(X_{n+1}|X_n) \ . \tag{2.10}$$

Du fait du modèle de Markov la quantité $h(X_{n+1}|X_n)$ a simplement pour expression $\sum_i \Pi_i^{(n)} h(x_{n+1}|x_n = i)$ où x_n désigne le n-ème caractère du texte X_{n+1} et $\Pi_i^{(n)}$ la probabilité inconditionnelle pour que x_n soit égal au symbole i. La quantité $h(x_{n+1}|x_n = i)$ est égale à $-\sum_j p_{ij} \log p_{ij}$, on obtient donc

$$h(X_n) = -\sum_{ij} (\Pi_i^{(0)} + \Pi_i^{(1)} + \cdots + \Pi_i^{(n-1)}) p_{ij} \log p_{ij}$$

Une expression simplifiée est possible avec le produit scalaire:

$$\langle (\Pi^{(0)} + \cdots + \Pi^{(n-1)}) \mathbf{L}, \eta \rangle \tag{2.11}$$

où $\Pi^{(n)}$ désigne le vecteur de distribution du n-ème caractère, $\mathbf{L}$ la matrice $p_{ij} \log p_{ij}$ et η le vecteur unitaire $(1, \ldots, 1)$. Comme $\Pi^{(n)} = \Pi^{(0)} \mathbf{P}^n$ où $\mathbf{P}$ désigne la matrice de transition p_{ij}, et que $\lim_n \Pi^{(n)} = \Pi$, Π désignant le vecteur de distribution stationnaire (Π_i), on obtient sans problème l'équivalence $h(X_n) \approx -n \langle \Pi \mathbf{L}, \eta \rangle$.

Il est possible de détailler davantage l'expression asymptotique de l'entropie markovienne en utilisant l'estimation $\langle \Pi^{(n)}, \eta \rangle = 1 + O(\theta^n)$ où $\theta < 1$ est la seconde valeur propre de la matrice $\mathbf{P}$. Il advient donc

$$h(X_n) = -n \langle \Pi \mathbf{L}, \eta \rangle + O(\theta^n) \ . \tag{2.12}$$

Noter que l'entropie d'un texte de Markov est différente de l'entropie H de l'état stationnaire d'une chaîne de Markov $H = -\sum_i \Pi_i \log \Pi_i$. Cette différence provient du fait que $h(X_n)$ mesure l'entropie de la chaîne de Markov complète, alors que H est relative à l'entropie du dernier maillon.

En passant nous prouvons que le modèle de Markov est mélangeant avec une densité d'entropie limite $h = -\sum_{ij} \Pi_i p_{ij} \log p_{ij}$.

2.3 Le calcul précis de l'entropie

2.3.1 Motivations

La théorie de l'information à ses débuts s'appliquait essentiellement au langage usuel. En l'absence de mathématisation du langage, il était inutile de s'attarder sur tel ou tel modèle plus ou moins markovien de l'information. En l'absence de modèles précis, les outils analytiques utilisés à cette époque étaient susceptibles de servir à tout type de modèles de sources. En conséquence les premières analyses asymptotiques des entropies étaient la plupart du temps limitées au premier ordre.

Néanmoins depuis la généralisation des échanges d'information entre ordinateurs, le calcul précis de l'entropie d'une source d'information ou de la capacité d'un canal a pris une importance qui dépasse les sphères de la théorie pure [5]. Les calculs ont dû dans certains cas être poussés au delà du premier ordre [7, 17] pour des modèles précis de sources.

En particulier on verra dans le chapitre suivant que les algorithmes de compression de données ont des performances qui sont limitées inférieurement par l'entropie des données à compresser. La distance qui sépare les performances effectives d'un algorithme de cette limite théorique s'appelle la *redondance* et sa détermination précise en fonction de la taille du texte à compresser permet de caractériser la qualité de l'algorithme. Un algorithme qui compresse efficacement à partir du centième symbole sera toujours plus intéressant qu'un algorithme qui doit attendre le dix-millième symbole.

La qualité des algorithmes de compression s'apprécie quand les tailles des textes à compresser sont importantes. En conséquence il y a lieu de déterminer les développement asymptotiques précis des entropies. Par exemple l'entropie d'un texte markovien de longueur n a un développement asymptotique calculé dans l'exercice commenté.

La détermination des développements asymptotiques de paramètres discrets est une branche importante de l'informatique théorique. En général il est pratique d'utiliser des méthodes analytiques sur des fonctions de variable complexe. Knuth [12] et Odlyzko [13] ont été des premier à insister sur le fait que "les méthodes analytiques sont extrêment puissantes et lorsqu'elles s'appliquent, elles apportent des évaluations d'une précision inégalées". La relation entre fonctions et coefficients discrets se traduit sous forme de séries génératrices [14, 16, 15, 17]. Dans ce contexte il est important de disposer de théorèmes de passages entre développements asymptotiques des séries génératrices et les développements asymptotiques des coefficients. C'est l'objet de la section consacrée à la dépoissonisation.

2.3.2 Les lèmmes de poissonisation

Les séries génératrices de Poisson sont des outils particulièrement adaptés pour la manipulation des variables aléatoires liées à la combinatoire, et de manière générale aux processus de Bernoulli ou de Markov. Elles sont donc indispensables à la théorie de l'information. Aldous [18] a montré que cette technique permet de dénouer efficacement les problèmes de dépendances croisées dans certains problèmes de combinatoire. Soit a_n une suite de coeffi-

cients donnée, la série de Poisson de cette suite a pour expression

$$f(z) = \sum_n a_n \frac{z^n}{n!} e^{-z}$$

Pour plus d'information sur les usages des séries de Poisson voir [23]. Dans cet article Jacquet et Szpankowski ont détaillé et généralisé les théorèmes de passage liés aux séries de Poisson.

Theorem 2 (Jacquet and Szpankowski [23]) *Soit un cône du plan complexe* $\mathcal{S}_\theta = \{z : |\arg(z)| \leq \theta, \ \theta < \pi/2\}$. *S'il existe une constante* β, $R > 0$ *et une fonction à variation lente* $L(x)$ *(pour tout* t $\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1$*) telles que les deux conditions suivantes sont satisfaites:*

- *(I) Pour* $z \in \mathcal{S}_\theta$*:*

$$|z| > R \quad \Rightarrow \quad |f(z)| \leq |z|^\beta \mathbf{\Lambda}(|z|) \ , \tag{2.13}$$

- *(O) Pour* $z \notin \mathcal{S}_\theta$*:*

$$|z| > R \quad \Rightarrow \quad |f(z)e^z| \leq M(|z|) \tag{2.14}$$

  où $M(x)$ *est une fonction qui décroit plus vite que n'importe quelle puissance de* x.

$$\begin{aligned} a_n &= \sum_{i=0}^{m} \sum_{j=0}^{i+m} b_{ij} n^i f^{\langle j \rangle}(n) + O(n^{\beta-m-1} L(n)) \qquad (2.15) \\ &= f(n) + \sum_{k=1}^{m} \sum_{i=1}^{k} b_{i,k+i} n^i f^{\langle k+i \rangle}(n) + O(n^{\beta-m-1} L(n)) \end{aligned}$$

où $f^{\langle j \rangle}(n)$ *est la* j*-ème derivée de* $f(z)$ *en* $z = n$, *et* b_{ij} *sont les coefficients de* $\exp(x \ln(1+y) - xy)$ *des facteurs* $x^i y^j$, *c'est à dire:*

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} b_{ij} x^i y^j = \exp(x \ln(1+y) - xy) \ . \tag{2.16}$$

tel que $b_{ij} = 0$ *pour* $j < 2i$. *En fait,* $b_{ij} = s_2(j,i)/j!$ *où* $s_2(n,k)$ *sont les nombres de Stirling associés de première espèce [22].*

Et le théorème inverse:

Theorem 3 *Soit* $g(z)$ *une fonction analytique telle que pour tout* n: $a_n = g(n)$. *On suppose qu'il existe* β *tel que* $g(z) = O(z^\beta)$ *pour* z *dans un cône complexe contenant le demi-axe réel positif. Alors il existe* θ *tel que dans le cône complexe* $\mathcal{S}_\theta$*:*

- *(I) pour tout* $z \in \mathcal{S}_\theta$: $|f(z)| \leq B|z|^\beta$ *pour une constante* $B > 0$;

– *(O) pour tout* $z \notin \mathcal{S}_\theta$*:* $|f(z)e^z| \leq Ae^{\alpha|z|}$ *pour* $A > 0$ *et* $\alpha < 1$.

De plus pour tout $w = O(z)$ *on a le développement asymptotique:*

$$\begin{aligned} f(z) &= \sum_{i=0}^{m} \sum_{j=0}^{i+m} a_{ij} z^i g^{\langle j \rangle}(z) + O(z^{\beta-m-1}) \\ &= g(z) + \sum_{k=1}^{m} \sum_{i=1}^{k} a_{i,k+i} z^i g^{\langle k+i \rangle}(z) + O(z^{\beta-m-1}) \end{aligned} \tag{2.17}$$

où a_{ij} *sont les coefficients de* $x^i y^j$ *dans* $\exp(x(e^y - 1) - y)$*. En fait,* $a_{ij} = S_2(j,i)/j!$ *où* $S_2(n,k)$ *sont les nombres de Stirling 2-associés de seconde espèce. En conséquence*

$$f^{\langle k \rangle}(z) = \sum_{i=0}^{m} \sum_{j=0}^{i+m} a_{ij}^{(k)} z^i g^{\langle j \rangle}(z) + O(z^{\beta-k-m-1}) \tag{2.18}$$

où $a_{ij}^{(k)}$ *est le coefficient de* $x^i y^j$ *dans* $(e^y - 1)^k \exp(x(e^y - 1) - y$.

2.3.3 Entropie de quelques distributions

Jacquet et Szpankowski [24] ont déterminé les développements asymptotiques des entropies pour une certaine classe de variables aléatoires: les variables de Bernoulli.

Ci-dessous nous donnons le développement asymptotique à un ordre arbitraire de l'entropie de la distribution binomiale $h_n = -\sum_k p_k^{(n)} \log p_k^{(n)}$ avec $p_k^{(n)} = \binom{n}{k} p^k q^{n-k}$ pour des constantes $p > 0$ et $q > 0$, $q = 1 - p$.

La distribution binomiale correspond à la distribution du nombre de caractères d'un type donnée dans une source de Bernoulli. Par exemple le nombre de consonnes dans un texte de longueur n, ou le nombre de lettres a. Dans la littérature [20] l'entropie de la distribution binomiale devait jusqu'à récemment se contenter de l'encadrement de premier ordre $\frac{1}{2}\log(\pi n/2) \leq h_n \leq \frac{1}{2}\log(\pi e n/2)$. Dans [21] il était montré que la borne supérieure est une limite asymptotiquement correcte jusqu'au deuxième ordre. Nous généralisons ce résultat jusqu'à un ordre arbitraire.

Theorem 4 *L'entropie* h_n *de la distribution binomiale possède le développement asymptotique suivant:*

$$\begin{aligned} h_n &= \frac{1}{2}\ln n + \frac{1}{2} + \frac{1}{2}\ln(2\pi pq) + \\ &+ \sum_{m=1}^{\infty} \frac{1}{n^{2m-1}} \left(\frac{B_{2m}(p^{1-2m} + q^{1-2m} - 1)}{2m(2m-1)} + o_m \right) + \sum_{m=1}^{\infty} \frac{e_m}{n^{2m}} \end{aligned} \tag{2.19}$$

où les B_k *sont les nombres de Bernoulli et* o_m *et* e_m *des constantes explicitement calculables à partir des constantes* b_{ij} *et* a_{ij} *des théorèmes 2 et 3.*

L'entropie de la distribution binomiale négative, c'est à dire lorsque $p_k^{(n)} = \binom{k}{n} p^{n+1} q^{k-n}$, est développable asymptotiquement de manière similaire. D'une manière générale les grandeurs de la forme

$$a_n = \sum_{k \geq n} \binom{k}{n} p^n q^{k-n} g(k) \tag{2.20}$$

où $g(z)$ est une fonction analytique sont développables grace aux théorèmes de passage de poissonisation et dépoissonisation. En ce qui concerne la distribution binomiale négative, on a $g(z) = \log(z!)$. Le résultat donne a_n exprimé comme une série double en $n^i g^{\langle j \rangle}(n/p)$ dont les coefficients sont explicites et l'ordre du terme d'erreur contrôlé. La méthodologie permet de traiter de manière fine les termes dits "add-β" mis en évidence dans [19].

2.4 Conclusion

La mesure de l'entropie est un élément fondamental de la théorie de l'information. Si l'entropie mesure la quantité d'information contenue dans un texte donné, elle ne mesure pas la *qualité* de l'information car cela fait partie de paramètres non quantifiables hors contexte. Le rôle de l'entropie se borne à mesurer la quantité d'informations différentes susceptibles d'être portées par un même support ou par des supports de la même famille. C'est donc une mesure du potentiel à la diversité d'une famille de supports plus qu'une mesure de l'information contenue dans une seule instance de support.

Par exemple l'entropie d'un monologue d'un bègue est la même que lorsqu'il est dit normalement, alors qu'il contient au moins deux fois plus de symboles. On mesure donc bien l'information, indépendamment de la méthode de traduction utilisée.

Le chapitre suivant traite de la compression de texte. La compression d'un texte conserve l'information, donc l'entropie des textes comprimés est la même que celles des textes originaux. L'entropie est donc un outil très utile pour analyser les performances des algorithmes de compression.

Bibliographie

[1] F. Pratt, *Secret and Urgent*, Blue Ribbon Books, 1939.

[2] H. Nyquist, "Certain factors affecting telegraph speed," *Bell Sys. Tech. J.*, vol 3, pp. 324-352, 1924.

[3] R. Hartley, "Transmission of Information," *Bell Sys. Tech. J.*, vol 7, pp. pp. 535-563, 1928.

[4] L. Boltzmann, "Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmertheorie und der Wahrscheilichkeitsrechnung respektive den Sätzen über das Wärmegleichgwicht," *Wien. Ber.*, vol 76, pp. 373-435, 1877.

[5] J. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Information and Control*, vol 18, pp. 203-219, 1971.

[6] A. Khinchin, "The entropy concept in probability theory," *Uspekhi Matematicheskikh Nauk.*, vol 8, pp. 3-20, 1953 (traduction anglaise dans *Mathematical Foundations of Information Theory*, Dover, 1957).

[7] C. Krattenhaler, P. Slater, "Asymptotic redundancy for universal quantum coding," preprint.

[8] W. Szpankowski, "On asymptotics on certain sums arising in universal coding," Problems of Information Transmissions, a paraître, 1998.

[9] C. Shannon, "Prediction and entropy of printed english," *Bell Sys. Tech. J.*, pp. 47-51, 1951.

[10] T. Cover, R. King, "A convergent gambling estimate of the entropy of English," *IEEE Trans. on Inform. Theory*, vol 24, pp. 413-421, 1978.

[11] L. Levitin, Z. Reingold, "Entropy of natural languages – theory and practice," *Chaos, Solitons and Fractals*, vol 4 pp. 709-743, 1994.

[12] D. Knuth, *The Art of Programming*, vol 1-3, Addison Wesley, 1973 et 1981.

[13] A. Odlyzko, "Asymptotic enumeration," *handbook of Combinatorics,* vol 2, Elsevier Science, pp. 1063-1229, 1995.

[14] P. Flajolet, A. Odlyzko, "Singularity analysis of generating functions," SIAM J. Discrete Methods 3, pp. 216-240, 1990.

[15] R. Sedgewick, P. Flajolet, *Analysis of Algorithms*, Addison Wesley, 1995.

[16] R. Sedgewick, P. Flajolet, *Analytic Combinatorics*, en préparation, INRIA TR-1888 (1993), TR-2026 (1993), TR-2376 (1994).

[17] W. Szpankowski, "On asymptotics on certain recurrences arising in universal coding," *Problems of Inf. Trans.*, 1998.

[18] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*, Springer Verlag, 1989.

[19] R. Krichevski, "Laplace's law of succession and universal coding," *IEEE Trans. on Inform. Theory*, vol 44, pp. 296-303, 1998.

[20] S. Chang, E. Weldon, "Coding for T-user multiple-access channels," *IEEE Trans. on Inform. Theory*, vol 25, pp. 684-691, 1979.

[21] B. Hughes, A. Cooper, "Near optimal multiuser codes for binary adder channel," *IEEE Trans. on Inform. Theory*, vol 42, pp. 387-398, 1996.

[22] L. Comtet, *Advanced Combinatorics*, Reidel Publishing, 1974.

[23] P. Jacquet and W. Szpankowski, "Analytical depoissonization and its applications," accepted in TCS, 1997

[24] P. Jacquet and W. Szpankowski, "Toward analytical information theory: entropy computations," submitted, 1997.

Chapitre 3

Compressions d'informations

Le bœuf qui voulait se faire grenouille

3.1 Introduction

Au départ le problème ressemble à la quadrature du cercle: comment faire tenir un texte de 100 lignes dans 30 lignes sans changer le contenu de l'information supportée? C'est un peu comme si on voulait vider une bouteille d'eau dans un dé à coudre, cela semble une gageure intenable. Et pourtant c'est possible. Et l'exploit est devenu tellement quotidien dans les échanges d'informations qu'il ne surprend plus personne.
Mais comment ça marche?

Essayons de procéder méthodiquement. Pour ce faire nous allons étudier un algorithme de compression de type "Bozzo". Prenons un texte de longueur n constitué de "0" et de "1". Il y a 2^n textes binaires de longueur n. Il y a aussi $2^n - 1$ textes de taille inférieure ou égale à $n-1$. Donc on peut mettre en correspondance presque tous les textes de taille n avec un texte un texte de taille inférieure ou égale à $n-1$; en fait tous, sauf 1 qui sera en correspondance avec lui même. L'algorithme de compression Bozzo 1.0 consiste à remplacer le texte initial par son correspondant. Si on suppose tous les textes initiaux comme équiprobables, on occupe en moyenne une place mémoire $n - 2 + (n+2)2^{-n}$ inférieure à la longueur initiale n.

Alors se pose la question inévitable: que se passe-t-il si on réitère le processus de compression sur le texte comprimé, et ainsi de suite. On a vu que sauf cas d'exception facilement évitable, la longueur du texte comprimé est inférieure d'au moins une unité à la longueur du texte initial. On arrivera donc à un texte sur-comprimé de longueur un en moins de n itérations. Mais qu'est-ce à dire de cet unique caractère qui permettrait de résumer tous les textes de longueur n. Il y a une erreur quelque part et il nous faut remonter jusqu'au début du processus.

Comment, après la première compression peut-on revenir en arrière et décomprimer le texte. Nous avons un texte de longueur $m < n$ et il suffit de regarder la table de correspon-

dance entre les textes de longueur n et les textes de longueur inférieure où égale à $n-1$. Mais comment savons nous que nous sommes en présence du texte initial de longueur n? si cela se trouve il s'agit d'un des textes intermédiaires de longueur plus petite et il conviendrait de consulter la table de correspondance pour cette longueur alternative.

Pour lever cette ambiguité, il faut que le texte compressé contienne aussi une information sur la longueur du texte initial, ce qui rajoute $\log_2 n$ bits. Cette information n'est pas directement liée au contenu du texte à compresser mais il est nécessaire à la décompression. Nous avons donc remplacer l'algorithme deffectueux Bozzo 1.0 par sa version améliorée Bozzo 1.1. Notons que le rajout des $\log_2 n$ symboles supplémentaires rend le texte compressé en général plus long que le texte initial et que probablement la famille des algorithmes Bozzo n'est peut être pas très bonne.

Mais il ne faut pas jeter la pierre trop vite, on verra plus loin que le modèle où tous les textes de longueur n sont équiprobables (en fait le modèle de Bernoulli uniforme) ne peut pas donner lieu à des compressions efficaces.

3.2 Contexte aléatoire de la compression

On a vu qu'il était impossible de compresser un texte de manière certaine et déterministe, au risque de se trouver devant le paradoxe de réduire tous les textes de taille n dans un seul caractère. Donc la compression d'un texte s'obtient dans un contexte aléatoire et en conséquence elle ne peut pas être définie en dehors des modèles probabilistes définis sur les textes.

Il est aussi important que les paramètres de l'algorithme pour reconstruire le texte original fassent partie integrante du code compressé. Certains auteurs, comme l'école de Kolmogorov [1, 2, 3], sont même allé plus loin en définissant la compression comme la création d'un code (binaire) susceptible de piloter un automate sur une machine de Türing afin de reconstituer le texte original. L'analyse de la complexité d'un tel automate montre des résultats analogues à ceux de l'analyse des performances de compression en théorie de l'information.

3.2.1 Le taux de compression bornée par l'entropie

On définit le taux de compression τ_n comme étant la moyenne divisée par n, des longueurs des codes comprimé des textes de taille n. Pour simplifier le discours, on suppose que le code comprimé est écrit dans le même alphabet que le texte original. Une compression est efficace dès qu'on a $\tau_n < 1$.

Theorem 5 *Dans l'hypothèse où le texte à compresser X_n et le code compressé $Z(X_n)$ sont écrits dans le même alphabet de taille V, on a l'inégalité:*

$$\tau_n \geq \frac{h(X_n)}{n \log V} \ . \tag{3.1}$$

Preuve: Comme la compression ne détruit, ni ne dénature l'information (on dit "compression sans perte"), et que tous les textes X_n de longueur n, peuvent être recouvrés, on a donc

$$h(Z(X_n)) = h(X_n) \tag{3.2}$$

En utilisant le théorème 1 on établit la borne:

$$h(Z(X_n)) \leq \mathrm{E}[|Z(X_n)|] \log V \tag{3.3}$$

où $|Y|$ indique la longueur d'un texte Y, et $\mathrm{E}[X]$ est l'espérance mathématique d'une variable aléatoire X. Avec $\mathrm{E}[|Z(X_n)|] = n\tau_n$, on retrouve l'inégalité annoncée.

Corollaire 1 *Les textes Bernoulli uniformes ne sont pas comprimables dans le même alphabet.*

3.2.2 Optimalité et redondance

Un algorithme de compression est dit optimal si le taux de compression est exactement égal à l'entropie. Kolmogorov [1] a le premier utilisé le terme "universel" pour qualifier des algorithmes de compression particulièrement intéressants. Un algorithme est dit universel si sa mise en œuvre ne nécessite pas de connaissance *a priori* sur la distribution de la source à compresser. On ne connaît pas d'algorithmes de compression qui soient universellement optimaux, indépendamment du modèle de texte. Par contre l'algorithme qui laisse tel quel le texte à compresser est optimal au regard du modèle de Bernoulli uniforme.

Un algorithme est asymptotiquement optimal si quelque soit le modèle de textes:

$$\lim_{n\to\infty} \frac{\tau_n n \log V}{h(X_n)} = 1 \tag{3.4}$$

Il existe de nombreux algorithmes asymptotiquement optimaux comme les algorithmes d'Huffman [4], Lempel-Ziv [9, 10], *etc.* En fait il est relativement facile de construire de tels algorithmes. Pour parvenir à pouvoir comparer ces algorithmes entre eux il importe de déterminer leurs redondances respectives.

On définit la redondance ρ_n d'un algorithme de compression par rapport à un modèle de texte:

$$\rho_n = \tau_n - \frac{h(X_n)}{n \log V} \tag{3.5}$$

Pour un algorithme asymptotiquement optimal on a $\lim \rho_n = 0$. En fait la vitesse de convergence de ρ_n vers zéro dépend du modèle de texte. Shields [8] a prouvé qu'il n'existe pas de vitesse de convergence universelle. Quelque soit l'algorithme de compression, quelque soit une suite arbitraire ε_n tendant vers zéro, il existe un modèle ergodique de texte dont la compression vérifie $\limsup(\rho_n/\varepsilon_n) = \infty$.

Dans ce qui suit on se propose de caractériser les redondances de deux célèbres algorithmes de Lempel et Ziv dont le caractère optimal a été montrer pour l'ensemble des sources ergodiques [5, 6, 7].

3.3 Les algorithmes Lempel et Ziv

Les algorithmes de compression de Lempel et Ziv ont été introduits en 1977 et 1978 par Ziv et Lempel. Ils ont la propriété d'être d'usage universel, c'est-à-dire qu'ils ne sont pas restreints à tel ou tel modèle de textes et ne nécessite pas de d'ajustements particulier pour passer d'un support à un autre. Ils donnent le taux de compression optimal quand la taille du texte à compresser croît vers l'infini.

Il existe deux algorithmes de compression Lempel et Ziv: l'algorithme de 1977 [9] et l'algorithme de 1978 [10]. Ils sont fonctionnellement très proches, l'algorithme de 1978 est un peu plus optimisé pour effectuer des compressions rapides, dites "en ligne".

3.3.1 L'algorithme de 1977

L'algorithme nécessite une base de données qui est un fragment du texte à comprimer. Pour amorcer l'algorithme on prend une entame du texte comme base de données. Il est important que la base de données contienne tous les symboles de l'alphabet.

L'algorithme procède de la manière suivante. Supposons qu'à l'étape m on ait déjà comprimé le texte jusqu'à la position $L_m = n$. La base de données est donc la portion du texte comprise dans l'intervalle $(1, n)$.

La nouvelle étape, l'étape $m+1$ consiste à déterminer à partir de la position n le plus grand segment de texte continu qui possède une copie dans la base de donnée. Soit ℓ_m la longueur de ce segment maximal et soit i la position de la copie de ce segment dans la base de donnée. D'après les notations de l'exercice commenté du premier chapitre on a $\ell_m = \max_j\{C_{jn}\} = C_{in}$.

On intègre ce nouveau segment à la base de données, ce qui revient à faire $L_{m+1} = L_m + \ell_m$ et dans le texte compressé le nouveau segment est remplacé par le couple (i, ℓ_m) qui indique que le segment doit être copié à partir de la position i sur une longueur ℓ_m.

Si on désire compresser le texte jusqu'à une position n donnée à l'avance, la condition stoppante de l'algorithme est $L_{m+1} \geq n$. Le nombre de segments obtenus est noté M_n

Dans le résultat final le texte compressé se présente par

1. une base de données initiale;
2. une succession de couples (pointeurs, longueurs).

3.3.2 Performances de l'algorithme 1977 en qualitatif

Examinons les performances de l'algorithme 1977 sur un texte de longueur n. Les pointeurs peuvent prendre indifféremment toutes les valeurs entre 1 et n. Il faut donc $\log_2 n$ bits pour les coder. Les longueurs des segments sont plus petites que la profondeur maximale du texte, qui est de l'ordre de $\log n$ si on se réfère au résultat de l'exercice commenté sur la profondeur d'un texte aléatoire. Donc il suffit de $\log\log n$ bits pour coder les longueurs.

Soit M_n le nombre de segments copiés. La taille donc du texte compressé est donc $B + (\log_2 n + O(\log\log n))M_n$ où B est la taille de la base de donnée initiale.

Pour approcher les performances de cet algorithme on définit la quantité $D_{j,n}$ par

$$D_{i,j} = \max_{i<n, i\neq j} \{C_{ij}\}$$

comme étant la profondeur du j-ème suffixe sur les n premiers suffixes. On a donc $\ell_m = D_{L_m,L_m}$ et par conséquent $L_{m+1} = D_{L_1,L_1} + \cdots + D_{L_m,L_m}$.

Quand le texte obéit au modèle de Bernoulli et quand on fixe n, Jacquet et Szpankowski [13] ont montré que la distribution de $D_{n,n}$ converge asymptotiquement vers la distribution $D^*_{n,n}$ que l'on obtiendrait si tous les suffixes étaient remplacés par autant de textes indépendants. La quantité $D^*_{n,n}$ étant la profondeur d'insertion dans un *trie* il suffit de se reporter à l'analyse des tries pour conclure que

$$\begin{aligned} \mathrm{E}[D_{n,n}] &= \frac{1}{h}(\log n + \gamma + \frac{h_2}{2h}) + P_1(\log n) + O(n^{-\varepsilon}) \ , \\ \mathrm{Var}(D_{n,n}) &= \frac{h_2 - h^2}{h^3} \log n + C + P_2(\log n) + O(n^{-\varepsilon}) \ , \end{aligned}$$

où h est l'entropie par symbole de l'alphabet (ou la densité d'entropie du modèle de Bernoulli), $h_2 = \sum_i p_i(\log p_i)^2$, γ la constante d'Euler et C une constante explicitement calculable. Les fonctions $P_1(x)$ et $P_2(x)$ sont périodiques avec de faibles amplitudes quand les quantités $\log p_i$ sont tous dans des rapports rationnels. Sinon les fonctions convergent simplement vers zéro lorsque x augmente.

Dans le cas uniforme $p_1 = p_2 = \cdots = p_V = 1/V$, la variance devient:

$$\mathrm{Var}(D_{n,n}) = \frac{\pi^2}{6\log^2 V} + \frac{1}{12} + P_2(\log n) + O(n^{-\varepsilon}) \ . \tag{3.6}$$

Dans le cas uniforme la distribution est asymptotiquement normale autour de sa moyenne et variance. Dans le cas uniforme la distribution converge de manière plus ponctuelle et on a une distribution limite *super-exponentielle*:

$$\lim_{n\to\infty} (\Pr\{D_{n,n} \leq x\} - \exp(-nV^{-x})) = 0 \tag{3.7}$$

Le résultat ci-dessus ne permettrait pas de conclure sur la distribution asymptotique de la variable aléatoire D_{L_m,L_m} lorsqu'elle est conditionnée par $L_m = n$. En effet cette variable aléatoire n'a pas la même distribution que $D_{n,n}$ parce que la condition $L_m = n$ introduit un biais dans la distribution de Bernoulli. Néanmoins par des arguments de croissance monotone on peut conclure que la longueur du dernier segment issu de la compression d'un texte de longueur n, c'est à dire D_{L_m,L_m} conditionné par $L_m < n$ et $L_{m+1} \geq n$ a pour moyenne asymptotique $\frac{\log n}{h} + O(1)$ et pour variance asymptotique $(\frac{h_2-h^2}{h^3}) \log n$ et est asymptotiquement normal autour de sa moyenne et variance (sauf dans le cas uniforme).

De la formule classique issue des processus à renouvellement,

$$\Pr\{M_n < m\} = \Pr\{L_m \geq n\} \tag{3.8}$$

on conclut que $\mathrm{E}[M_n]$ est asymptotiquement équivalent à $(\frac{h}{\log n} + O(1))n$, donc que la taille du code compressé est en moyenne de $(\frac{h}{\log V} + O(\frac{\log\log n}{\log n}))n$.

Ce qui permet de conclure que l'algorithme Lempel-Ziv 77 est asymptotiquement optimal sur les textes de Bernoulli et sa redondance est en $O(\frac{\log\log n}{\log n})$.

3.3.3 L'algorithme 1978

L'algorithme ne nécessite pas de base de données initiale. Comme l'algorithme précédent il consiste à diviser le texte initial en phrases successives. Mais il faut maintenant que toute nouvelle phrase soit la plus grande phrase qui puisse être constituée d'une des phrases stockées dans la base de données augmentée d'un symbole.

L'opération diffère sensiblement de celle de l'algorithme de 1977 où chaque nouvelle phrase doit être identique à un segment dans la base de données précédente même si ce segment se trouve à cheval sur plusieurs phrases.

La nouvelle phrase est intégrée dans la base de données sous la forme d'un segment de texte, adjoint d'un pointeur sur une phrase précédente plus un identificateur du symbole supplémentaire. La première phrase est limité à un seul caractère et pointe sur une phrase symbolique vide.

On appelle I_m la longueur de la nouvelle phrase détectée à l'étape m. On appelle L_m la longueur cumulée des phrases précédentes: $L_m = I_1 + \cdots + I_{m-1}$. La condition stoppante de l'algorithme est $L_{m+1} \geq n$ si l'on désire compresser un texte jusqu'à la position n.

Par exemple le texte 110010100010001000 est découpé en l'ensemble des phrases:

$$(1)(10)(0)(101)(00)(01)(000)(100) \tag{3.9}$$

et la liste des couples (pointeurs, symboles) est

$$([0, \text{`1'}], [1, \text{`0'}], [0, \text{`0'}], [2, \text{`1'}], [3, \text{`0'}], [3, \text{`1'}], [5, \text{`0'}], [2, \text{`0'}])\ . \tag{3.10}$$

Le code compressé est égal à la listes des couples (pointeurs, symboles). Si M_n est le nombre de phrases obtenues en découpant le texte jusqu'à la position n, le code compressé pèse: $M_n \log_2 n + \log_2 V$.

La recherche du pointeur associé à la m-ème phrase peut être avantageusement effectué à l'aide d'un arbre digital de recherche qui rassemble les informations pertinentes sur les phrases précédentes. Pour une description détaillée de l'arbre digital de recherche voir la section consacrée aux structures de données. Il suffit de savoir que la quantité L_m est exactement la longueur de cheminement externe de l'arbre digital associé.

Jacquet et Szpankowski [14] ont montré que sur des textes de Bernoulli le paramètre M_n satisfait:

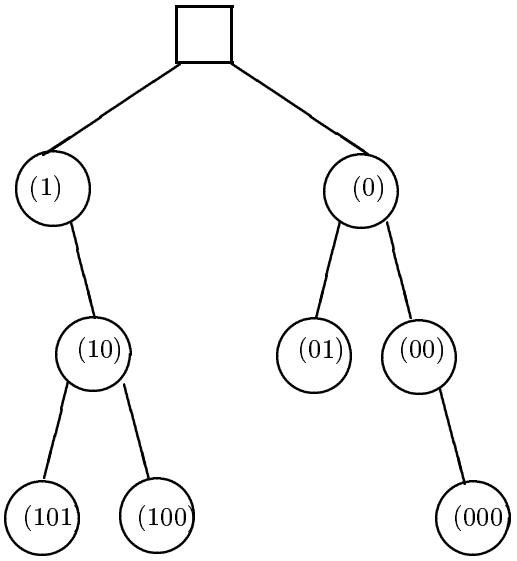

FIG. 3.1 – *L'arbre digital associé au découpage du texte* 11001010001000100...

$$\begin{aligned} \mathrm{E}M_n &= \frac{nh}{\log n + \frac{h_2}{2h} + \gamma - 1 - \alpha + \delta_0(\log n)} + o(n) \\ \mathrm{Var}(M_n) &= (\frac{h_2 - h^2}{h^3})\frac{n}{\log n} + O(\frac{n}{\log^2 n}) \end{aligned}$$

la quantité h étant l'entropie de l'alphabet et h_2 l'entropie seconde, c'est-à-dire $\sum_i p_i \log^2 p_i$, $\alpha = -\sum_k (\sum_i p_i^{k+1} \log p_i)/(1 - \sum_i p_i^{k+1}$. La fonction $\delta_0(x)$ est une fonction périodique de faible amplitude quand les $\log p_i$ sont tous dans des proportions rationnelles, ou égale à zéro autrement.

Quand le modèle est Bernoulli uniforme, le facteur de $n/\log n$ dans $\mathrm{Var}(M_n)$ s'annule, par contre le facteur de $n/\log^2 n$ est de la forme $C + \delta_1(\log n) + o(n)$ où $\delta_1(x)$ vérifie les mêmes propriétés que $\delta_0(x)$.

De plus la variable aléatoire M_n est asymptotiquement normale sur sa moyenne et variance.

L'analyse utilise l'identité de renouvellement (3.8) et une analyse fine de la distribution limite de la longueur de cheminement dans un arbre digital.

En conséquence la redondance ρ_n est asymptotiquement normale avec moyenne $(\alpha + 1 - \gamma - \frac{h_2}{2h} - \delta_0(\log n) + o(n))(\log n)^{-1}$ et variance $O((\sqrt{n}\log n)^{-1})$. Donc la redondance de l'algorithme 1978 sur les textes Bernoulli est en $O(\frac{1}{\log n})$. Avant cette étude la moyenne de

la redondance était conjecturée en $O(\frac{\log \log n}{\log n})$ comme celle de l'algorithme 1977. Ce résultat est un peu surprenant, l'intuition donnant plutôt l'algorithme 1977 gagnant dans la mesure où la recherche des phrases dupliquées y est plus exhaustive.

3.3.4 Conclusion

La compression d'information est devenue un outil d'importance cruciale pour les télécommunication et le stockage d'informations. Par ailleurs de nouvelles applications surprenantes sont actuellement mise en œuvre. Par exemple, les linguistes utilisent les algorithmes universels de compression Lempel et Ziv pour déterminer l'entropie des langues usuelles [11, 12]. Une autre application consiste à comprimer un texte d'origine inconnue sur des bases de données de langues différentes afin de déterminer l'origine linguistique du texte sans analyse sémantique (voir figure 3.2). La base de donnée qui donnera la compression minimale correspondra à l'origine linguistique recherchée. Par une méthode équivalente, mais plus fine, on peut déterminer l'auteur d'un texte anonyme (voir figure 3.3).

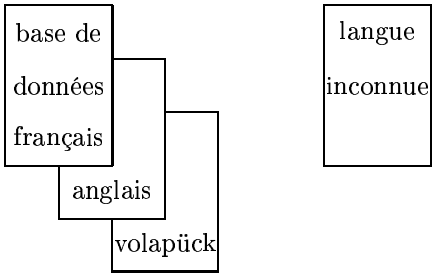

FIG. 3.2 – *Application de la compression de texte à l'analyse linguistique*

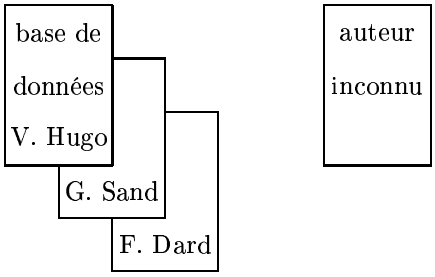

FIG. 3.3 – *Application de la compression à l'analyse littéraire*

Bibliographie

[1] A. KOLMOGOROV, "Three approaches to the quantitative definition of information," *Problems of Information Transmisssion*, vol 1, pp. 1-7, 1965.

[2] A. KOLMOGOROV, "Logical basis for information theory and probability theory," *IEEE Trans. Inform. Theory*, vol 14, pp. 662-664, 1968.

[3] R. SOLOMONOFF, "A formal theory of inductive inference," *Information and Control*, vol 7, pp. 1-22, 224-254, 1964.

[4] D. HUFFMAN, "A method for the construction of minimum redundancy codes," *Proc. IRE*, pp. 1098-1101, 1952.

[5] J. ZIV, "Coding theorems for individual sequences," *IEEE Trans. Inform. Theory*, vol 24, pp. 405-412, 1978.

[6] T, COVER, J. THOMAS *Elements of Information Theory*, Wiley, 1991.

[7] A. WYNER, J. ZIV "The sliding-window Lempel-Ziv algorithm is asymptotically optimal," *Proc. IEEE*, pp. 872-877, 1994.

[8] P. SHIELDS, "Universal Redundancy Rates Do Not Exist," *IEEE Information Theory*, vol 39, 520-524, 1993.

[9] J. ZIV, A. LEMPEL, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol 23, pp. 337-343, 1977.

[10] J. ZIV, A. LEMPEL, "Compression of individual sequences via variable-rate coding, *IEEE Trans. Inform. Theory*, vol 24, pp. 530-536, 1978.

[11] I. KONTOYIANNIS, P. ALGOET, Y. SUHOV, A. WYNER, "Non-parametric entropy estimation for stationnary process and random fields, with application to English text," *IEEE Trans. Inform. Theory*, vol 44, pp. 1319-1327, 1998.

[12] A. WYNER, J. ZIV, A. J. WYNER, "On the role of pattern matching in information theory," *IEEE Trans. Inform. Theory*, to appear.

[13] P. JACQUET, W. SZPANKOWSKI, "Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach," in *J. Combinatorial Theory. Ser. A.* May 1992.

[14] P. JACQUET AND W. SZPANKOWSKI, "Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees," Theoretical Computer Science, volume 144, pp. 161-198, juin 1995. *Nominé pour Best Paper Award 1997.*

Chapitre 4

Les structures de données

Depuis Adam et Ève, en passant par Sir Isaac Newton, la connaissance nous est souvent tombée des arbres

4.1 Introduction

Une structure de données est associée à une mémoire pour stocker un nombre arbitraire n de segments d'informations et à un automate pour accéder à n'importe lequel de ces n segments. L'automate prend pour paramètre d'entrée un identifiant du segment à consulter, appelé *clef d'accès* et rend en sortie un pointeur sur la mémoire qui contient le segment d'information recherché.

La clef d'accès est une fonction fixée à l'avance qui prend en variable une portion des données présentes sur le segment à recouvrer et seulement sur celui-ci. En d'autres termes la clef d'accès ne dépend que de la donnée à accéder et ne nécessite pas de connaissance sur les autres données stockées en parallèle. En général la clef d'accès est constituée des premiers caractères significatifs du segment d'information à consulter. Par exemple la clef d'accès à la biographie d'Isaac Newton dans la structure de données de l'*Encyclopædia Britannica* est dans les premiers caractères de "Newton, Sir Isaac".

L'automate de la structure de données sert aussi à l'insertion éventuelle de nouveaux segments d'information dans la structure. Les premiers travaux sur l'évaluation systématique des algorithmes des structures de données datent de Knuth [1], [2], [3].

Le nombre d'opérations effectuées par l'automate pour accéder à un segment d'information donné s'appelle la *profondeur d'insertion* dudit segment. Cette profondeur d'insertion a les caractéristiques d'une variable aléatoire quand la sélection du segment à recouvrer est elle-même aléatoire ou, mieux, quand les n segments stockés sont eux aussi issus d'une source d'information aléatoire.

Étant donné n segments d'information stockés, le nombre minimum d'opérations pour accéder à l'un d'entre eux, c'est à dire pour le discriminer par rapport aux $n-1$ autres

segments, peut être déterminé en référence à la théorie de l'information. Une suppose que la taille V de l'alphabet coincide avec le degré de l'automate à chacune de ses étapes, c'est à dire le nombre de cas possibles à chaque opération, la profondeur d'insertion minimale théorique est asymptotiquement équivalente à

$$\frac{\log n}{h} \ , \tag{4.1}$$

où h est la densité d'entropie si elle existe de la source d'information:

$$h = \lim_{n \to \infty} (h(X_{n+1}) - h(X_n))$$

Pour parvenir à ce résultat il suffit d'imaginer le nombre de caractères qu'il faut extraire de chaque segment pour les discriminer de façon certaine. Soit k le nombre de caractères extraits, du fait de l'hypothèse de la densité d'entropie, l'entropie de ces k caractères a pour valeur asymptotique $k.h$. L'entropie de l'information nécessaire pour distinguer n éléments entre eux est au moins $\log n$, ou $\log_2 n$ en bit, ce qui est plus parlant (il faut $\log_2 n$ bits pour coder de manière distinctes n éléments). On arrive donc à l'inégalité $k.h \geq \log n$.

4.2 Analyse des tries

Le *trie* est une structure de données qui suit la forme d'un arbre. Les données sont consultables à partir des feuilles. Dans la mesure où les données sont des textes écrits dans un alphabet de taille V, chaque nœud interne de l'arbre est racine d'un branchement de degré V. À chaque symbole de l'alphabet correspond une arête dans l'embranchement qui part de chaque nœud interne. Pour des données binaires, le *trie* est donc un arbre binaire.

Pour accéder à la feuille qui contient la donnée que l'on désire recouvrer il suffit de parcourir l'arbre à partir de sa racine en choisissant les arêtes successives conformément aux premiers caractères de ladite donnée.

En conséquence les premiers caractères de chaque donnée jouent le rôle de clef d'accès au *trie*. La longueur du préfixe utilisé comme clef est donc la profondeur d'insertion de la donnée. Par définition c'est le nombre d'embranchements sur les nœuds internes que l'on doit traverser à partir de la racine du *trie* afin d'accéder à la feuille de référence de la donnée.

La profondeur peut être plus ou moins longue en fonction de la contenance maximale des feuilles. Si la contenance de chaque feuille est limitée à une seule donnée à la fois, la profondeur d'insertion sera nécessairement plus longue que dans les cas de contenance supérieure.

Le *trie* peut être constitué par insertion successive des données. L'arbre final ne dépend pas de l'ordre d'insertion. Si la clef d'accès d'une nouvelle donnée implique un cheminement dans l'arbre qui aboutit dans une feuille et que cette feuille a déjà atteint sa contenance maximale, alors la feuille est remplacé par un nœud interne et son contenu est distribué sur les branchements ainsi crées.

L'arbre suffixe est un cas particulier du *trie* où les feuilles sont de contenance un et les données sont les n premiers suffixes d'un texte de référence. La profondeur d'un suffixe telle

qu'on l'a définie au chapitre sur la compression est la profondeur d'insertion du suffixe dans l'arbre suffixe.

Les *tries* ont été analysés sous des modèles de Bernoulli sur alphabet binaire, en moyenne par Flajolet [4], au second moment par Jacquet et Régnier [5], sur un alphabet V-aire par Szpankowski [6]. Kirschenhofer et Prodinger [7] ont montré des développements asymptotiques similaires sur les arbres digitaux. Pittel [8] a montré que la profondeur d'insertion dans le *trie* divisé par $\log n$ et l'entropie converge vers 1 en probabilité dans les modèles mélangeants.

4.2.1 Le modèle de Bernoulli

Jacquet et Régnier [9] ont calculé les développements asymptotiques des moyennes et variances de la taille du *trie* dans le cas de données obéissant au modèle de Bernoulli. Ils ont aussi démontrer la normalité de la loi limite de la taille du *trie* quand le nombre d'insertions n tend vers l'infini. Ils ont aussi calculé les développements asymptotiques de la moyenne et de la variance de la profondeur d'insertion D_n quand on choisit au hasard une donnée à consulter. Ils ont caractérisé la distribution limite (voir le chapitre consacré à la compression).

La distribution de D_n quand les feuilles peuvent contenir b données, vérifie l'équation fonctionnelle:

$$D(z,u) = u\sum_i p_i D(p_i z, u) + (1-u)(1 + z + \cdots + \frac{z^{b-1}}{(b-1)!})e^{-z} \tag{4.2}$$

où $D(z,u)$ est la transformée de Poisson de $D_n(u) = \sum_k \Pr\{D_n = k\}u^k$. On a

$$\begin{aligned} \mathrm{E}[D_n] &= \frac{\partial}{\partial u} D_n(1) \\ \mathrm{Var}(D_n) &= \frac{\partial^2}{\partial u^2} D_n(1) + \frac{\partial}{\partial u} D_n(1) - (\frac{\partial}{\partial u} D_n(1))^2 \end{aligned}$$

En utilisant la transformée de Mellin et la dépoissonisation, on obtient avec contenance $b = 1$:

$$\begin{aligned} \mathrm{E}[D_n] &= \frac{1}{h}(\log n + \gamma + \frac{h_2}{2h} - h) + P_1(\log n) + O(n^{-\varepsilon}) \ , \\ \mathrm{Var}(D_n) &= \frac{h_2 - h^2}{h^3}\log n + C + P_2(\log n) + O(n^{-\varepsilon}) \ , \end{aligned}$$

On remarque donc que le *trie* présente des performances asymptotiquement optimales pour le modèle de Bernoulli. Dans la section suivante, on montre que c'est aussi le cas pour le modèle de Markov ce qui confirme bien les résultats dans [8]. De plus les résultats asymptotiques sur la distribution limite peuvent être poussés à des ordres arbitraires.

4.2.2 Le modèle de Markov

Jacquet et Szpankowski [10] ont calculé les développements asymptotiques de la profondeur d'insertion quand les données obéissent à un modèle de Markov. Dans ce cas on retrouve pratiquement les mêmes expressions que dans l'identité (4.3) mais avec quelques modifications

$$\begin{aligned} \mathrm{E}[D_n] &= \frac{1}{h}(\log n + \gamma + \frac{h_2}{2h} - H) + P_1(\log n) + O(n^{-\varepsilon}) \ , \\ \mathrm{Var}(D_n) &= \frac{h_2 - h^2}{h^3} \log n + C + P_2(\log n) + O(n^{-\varepsilon}) \ . \end{aligned}$$

La quantité h est l'entropie de l'alphabet telle que déterminée dans le chapitre consacré aux calculs d'entropies: $h = -\sum_{ij} \Pi_i p_{ij} \log p_{ij}$, la quantité $H = -\sum_i \Pi_i \log \Pi_i$. La quantité h_2 est la dérivée seconde au point $t = 1$ de la valeur propre principale $\lambda(t)$ de la matrice $\mathbf{P}(t) = [p_{ij}^t]$: $h_2 = \ddot{\lambda}(1)$. Il existe une expression un peu compliquée basée sur les dérivées des matrices $\mathbf{P}(t)$, de ses vecteurs propres gauches $\Pi(t)$ et droits $\eta(t)$:

$$h_2 = \ddot{\lambda}(1) = \langle \Pi(1)\ddot{\mathbf{P}}(1), \eta(1) \rangle - \langle \dot{\Pi}(1)\dot{\mathbf{P}}(1), \eta(1) \rangle - \langle \Pi(1)\dot{\mathbf{P}}(1), \dot{\eta}(1) \rangle \tag{4.3}$$

Noter qu'on a également

$$h = \dot{\lambda}(1) = \langle \Pi(1)\dot{\mathbf{P}}(1), \eta(1) \rangle \ . \tag{4.4}$$

En outre la distribution limite est asymptotiquement normale (exceptée l'exception du cas uniforme du modèle Bernoulli).

Un dernier point: la condition pour avoir les fonctions $P_1(x)$ et $P_2(x)$ périodique est légèrement différente de celle du cas Bernoulli. En effet Tang dans sa thèse de PhD a démontré que cette condition est atteinte quand les quantités $\log p_{ij} + \log p_{1i} - \log p_{1j}$ sont dans des proportions rationnelles. Un modèle de Markov dégénéré (donc un modèle de Bernoulli) où $p_{ij} = p_i$ pour tout i et j et tel que tous les $\log p_i$ sont en proportions rationnelles vérifie trivialement cette condition.

Jacquet et Szpankowski ont montré qu'il existe des modèles de Markov qui ne se réduisent pas à des modèles de Bernoulli (ou modèle de Markov dégénéré) et qui vérifient quand même la condition de périodicité. Par exemple, soit une matrice $[k_{ij}]$ d'entiers arbitraires et soit $\mathbf{M}(t)$ la matrice $[e^{-2\pi k_{ij}/t}]$: on dénote par $\lambda(t)$ sa valeur propre principale.

Il est clair que $\lambda(0) = 0$ et $\lim_{t\to\infty} \lambda(t) = V$. Il existe t_0 tel que $\lambda(t_0) = 1$. Soit $r_i(t)$ les composantes du vecteur propre droit de $\mathbf{M}(t)$. On définit la matrice $\mathbf{P} = [p_{ij}]$ par l'identité

$$p_{ij} = \frac{r_j(t_0)}{r_i(t_0)} e^{-2\pi k_{ij}/t_0} \tag{4.5}$$

La matrice $\mathbf{P}$ ainsi définie est une matrice de Markov puisque pour tout i:

$$\sum_j p_{ij} = \frac{1}{r_i(t_0)} \sum_j r_j(t_0) e^{-2\pi k_{ij}/t_0} = \frac{r_i(t_0)}{r_i(t_0)} = 1 \tag{4.6}$$

La matrice $\mathbf{P}$ ne correspond pas au modèle dégénéré puisque les k_{ij} sont arbitraires. La vérification de la condition de périodicité est laissée en exercice.

4.3 L'arbre digital de recherche

L'arbre digital de recherche (DST) a une structure d'arbre comme le *trie*. Chaque nœud interne est de degré inférieur ou égal à la taille de l'alphabet, et à chaque branchement les arêtes correspondent à des symboles de l'alphabet. À la différence du *trie* le DST porte ses données dans ses nœuds internes. Comme dans le *trie* le chemin d'accès à chacune des données stockées dans les nœuds correspond à un préfixe de ladite données. Le DST dépend de l'ordre d'insertion des données. Lors de l'insertion d'une nouvelle donnée, on parcourt l'arbre à partir de la racine en suivant un chemin dont les arêtes correspondent dans l'ordre aux premiers caractères de la donnée. On stocke la nouvelle donnée dans le premier nœud visité qui n'a pas encore atteint sa contenance maximale. Si tous les nœuds visités ont déjà atteint leur contenance maximale, alors un nouveau nœud est créé sous le dernier nœud visité pour recevoir la donnée ; la nouvelle arête ainsi créée est indexée par le symbole du caractère de la donnée qui permet d'identifier le chemin jusqu'à ce nouveau nœud.

La distribution de la profondeur d'insertion dans un DST est connue. Tang et Szpankowski ont montré que dans le modèle de Markov la profondeur d'insertion D_n a une distribution limite normale et

$$\begin{aligned} \mathrm{E}[D_n] &= \frac{1}{h}(\log n + \gamma + \frac{h_2}{2h} - H + C_1) + Q_1(\log n) + O(\frac{\log n}{n}) \ , \\ \mathrm{Var}(D_n) &= \frac{1}{h^3}(C_2 - h^2)\log n + O(1) \ , \end{aligned}$$

où C_1 et C_2 sont des constantes calculables explicitement en fonction de la matrice $\mathbf{P}(s)$ et de ses dérivées, les expressions sont un peu compliquées et sont donc omises. La fonction $Q_1(x)$ tend vers zéro quand $x \to \infty$ excepté dans les conditions de périodicité identiques à celle du *trie*.

Nous avons vu dans le chapitre sur la compression que la détermination de la redondance de l'algorithme Lempel et Ziv 78 passe par l'analyse asymptotique de la longueur de cheminement externe L_n dans le DST. La longueur de cheminement externe est, rappelons le, la somme des profondeurs d'insertion des n données stockées dans le DST. Dans le modèle de Bernoulli et pour des feuilles de DST de contenance 1, Jacquet et Szpankowski [11] ont montré que la distribution de L_n est asymptotiquement normale et:

$$\begin{aligned} \mathrm{E}[L_n] &= \frac{n+1}{h}(\log n + \frac{h_2}{h} + \gamma - 1 - \alpha + \delta_0(\log n)) + O(1) \\ \mathrm{Var}(L_n) &= \frac{(h_2 - h^2)}{h^3} n \log n + O(n) \end{aligned}$$

où

$$\alpha = -\sum_{k=1}^{\infty} \frac{\sum_i p_i^{k+1} \log p_i}{1 - \sum_i p_i^{k+1}} \ , \tag{4.7}$$

dans lequel $\delta_0(x)$ est une fonction qui tend vers zéro quand $x \to \infty$ à l'exception des cas de périodicité déjà évoqués dans l'analyse des *tries*. On montre aussi un résultat de convergence dominée: pour tout θ dans un voisinage complexe de zéro et pour tout $\varepsilon > 0$

$$\exp(-\theta \frac{n}{h} \log n) \mathrm{E}[e^{\theta L_n}] = \exp(c_2 \frac{\theta^2}{2} n \log n)(1 + O(m^{-1/2+\varepsilon})) \ . \tag{4.8}$$

L'analyse s'appuie sur la dépoissonisation et sur l'analyse asymptotique de l'équation différentielle non linéaire aux différences, un peu délicate à traiter:

$$\frac{\partial^b}{\partial z^b} L(z,u) = \prod_i L(p_i z u, u) \tag{4.9}$$

où b est la contenance des nœuds du DST et

$$L(z,u) = \sum_n \frac{z^n}{n!} \mathrm{E}[u^{L_n}] \ . \tag{4.10}$$

Le résultat central est l'évaluation pour z variant dans un cône convexe du plan complexe (voir figure **??**).

$$\log(L(z,u) = O(z^{\kappa(u)}) \ , \tag{4.11}$$

avec $\kappa(u)$ déterminé algèbriquement par l'identité:

$$(\sum_i p_i^{\kappa(u)}) u^{\kappa(u)} = 1 \ . \tag{4.12}$$

L'évaluation asymptotique de $\log(L(z,u)$ permet de prouver de manière analytique la normalité de la loi limite de la longueur de cheminement externe.

Des résultats similaires sont accessibles pour des valeurs de b arbitraires et dans le modèle de Markov mais ils n'ont pas été développés par manque de place. Sous le modèle de Markov on a alors l'identité

$$\det(u^{\kappa(u)} \mathbf{P}(\kappa(u)) - I) = 0 \ , \tag{4.13}$$

où $\det(M)$ désigne le déterminant d'une matrice $V \times V$ M et I désigne la matrice identité.

4.4 Conclusion

Les structures de données constituent un domaine à l'intersection de la théorie de l'information et le domaine des algorithmes. C'est aussi un domaine qui a été un des plus riches en applications, notamment sur le plan du génie logiciel, des bases de données et des télécommunications. Il n'est donc pas étonnant qu'il donne lieu à des développements analytiques très avancés depuis de nombreuses années [1].

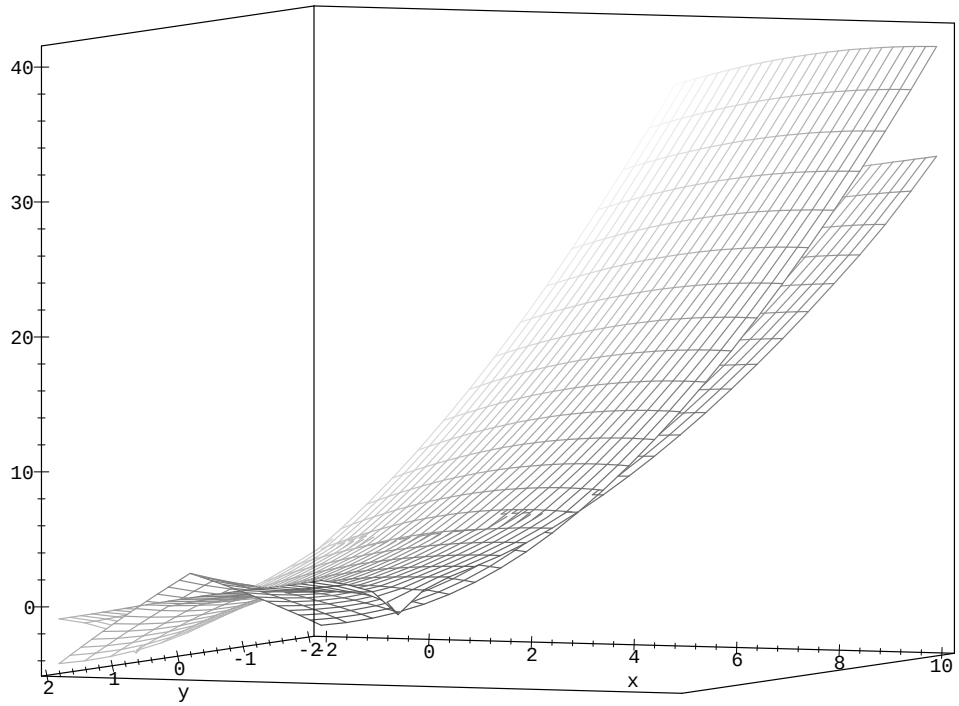

FIG. 4.1 – $\log(L(z,u))$ *et son asymptotique* $O(z^{\log(2)/\log(2/u)})$ *dans le plan complexe* ($u = 1.3$)

Bibliographie

[1] D. KNUTH, *The Art of Computer Programming, Sorting and Searching*, Addison-Wesley, 1973.

[2] R. FAGIN, J. NIEVERGELT, N. PIPPENGER, H. STRONG, "Extendible hashing: a fast access method for dynamic files," *ACM TODS*, vol 4, pp. 314-344, 1979.

[3] A. AHAO, J. HOPCROFT, J. ULLMAN, *Data Structures and Algorithms*, Addison-Wesley, 1983.

[4] P. FLAJOLET, "On the performance evaluation of extendible hashing and trie searching," *Acta Informatica*, vol 20, pp. 345-369, 1983.

[5] M. RÉGNIER, P. JACQUET, "New results on the size of tries," *IEEE Trans. Inform. Theory*, vol-35, pp. 203-205, 1989.

[6] W. SZPANKOWSKI, "Some results on V-ary asymmetric tries," *J. Algorithms*, vol 9, pp. 224-244, 1988.

[7] P. KIRSCHENHOFER, H. PRODINGER, "Further results on digital search trees," *Theoretical Comput. Science*, vol 58, pp. 143-154.

[8] B. PITTEL

[9] P. JACQUET, M. RÉGNIER, "Trie partioning process: limiting distributions," *Lecture Notes in Computer Science*, vol 214, pp. 196-210, Springer, New York, 1986.

[10] P. JACQUET, W. SZPANKOWSKI, "Analysis of digital tries with Markovian dependency," *IEEE Trans. Inform. Theory*, vol-37, pp. 1470-1475, 1991.

[11] P. JACQUET AND W. SZPANKOWSKI, "Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees," Theoretical Computer Science, volume 144, pp. 161-198, juin 1995.

Chapitre 5

Protocoles de communications

Et Dieu créa Internet. À moins que ça ne devienne le contraire...

5.1 Introduction

Dans le chapitre précédent nous avons regardé les structures de données aptes à stocker et recouvrir de l'information. Dans ce chapitre nous examinons plus particulièrement les moyens pour faire circuler l'information. Ces moyens sont principalement les réseaux de télécommunication.

L'énorme impact de la généralisation d'Internet donne un relief plus accentué à ce domaine. Nous allons donner deux exemples d'analyses issues de l'algorithmique et de la théorie analytique de l'information. Le premier est plus proche du domaine applicatif, il concerne l'analyse des algorithmes de résolution de collisions qui ont été proposés et adoptés comme standard pour les réseaux de modems câble pour l'accès à internet à hauts débits. Le second, plus théorique, concerne l'analyse des corrélations à long terme que l'on observe actuellement dans le trafic internet. L'analyse repose sur des superpositions de sources *on/off* poissonniennes.

Les dénominateurs communs à la plupart des réseaux de données actuels sont

- le transfert d'informations sous forme de segments multiples et séparés, les *paquets*;
- le multiplexage des sources sur des canaux à diffusion ou à accès multiples.

Un canal à diffusion est un canal qui peut supporter simultanément des communications différentes destinées à des récepteurs différents (voir figure 5.1. Un canal à accès multiples est un canal qui peut supporter simultanément des communications différentes en provenance d'émetteurs différents (voir figure 5.2. Les réseaux Ethernet et, à plus haut niveau, le réseau Internet sont à diffusion et à accès multiples. Historiquement, le premier canal physique à diffusion et à accès multiples a été inventé par Edison et Bell qui eurent l'idée en 1870 de

transmettre des signaux télégraphiques simultanés sur un seul fil. Les canaux à diffusion ou à accès multiples ont été décrits de manières théoriques dans [3] (voir figure 5.3.

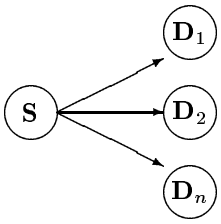

FIG. 5.1 – *Canal à diffusion*

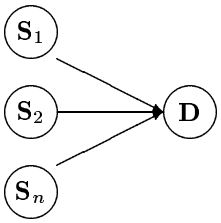

FIG. 5.2 – *Canal à accès multiple*

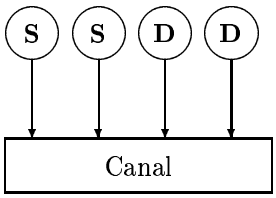

FIG. 5.3 – *Canal à diffusion et à accès multiple*

5.1.1 Les capacités d'un canal de communication

Capacité maximale théorique et capacité réelle

On a vu dans le chapitre consacrée à l'entropie de l'information que la capacité maximale C_m d'un canal de communication était donnée par la formule

$$C_m = \max_X \{h(T(X)) - h(T(X)|X)\} \tag{5.1}$$

FIG. 5.4 – *Format des paquets transitant sur un réseau*

où X est le segment d'information envoyé par la source et $T(X)$ le segment décodé par le récepteur.

Cette capacité ne tient pas compte des erreurs de transmissions et autres distorsion dues au canal. Il se peut par exemple que le canal dénature complètement l'information transmise. Pour parvenir à ce que le récepteur recouvre exactement l'information transmise par l'émetteur il faut introduire le concept du codage et du décodage:

1. L'émetteur code son fragment d'information X_n: $C(X_n)$;
2. L'émetteur transmet $C(X_n)$, et le récepteur reçoit $Y_n = T(C(X_n))$;
3. le récepteur décode Y_n: $D(Y_n)$.

Il y a erreur de transmission quant $D(Y_n) \neq X_n$.

Le grand théorème de Shannon [1] démontre que C_m est atteignable avec une probabilité d'erreur arbitrairement faible. On a une erreur lorsque le décodage du code reçu fournit un segment d'information différent du segment original. Si on appelle T^{-1} la fonction de décodage la probabilité d'erreur est donc $\Pr\{T^{-1}(T(X)) \neq X\}$.

Theorem 6 (Grand théorème de Shannon) *Soit un canal donné tel que $C_m > 0$. Pour tout $\varepsilon > 0$ et pour tout $C < C_m$ il existe un système de codage et de décodage de l'information tel que le taux d'erreur est inférieur à ε et la quantité d'information transmise est supérieure à C.*

La capacité C_m est atteignable dans la mesure où l'on dispose d'un code correcteur d'erreur universellement efficace. Ces codes n'existent pas dans la pratique et leur existence n'est envisageable que dans des perspectives asymptotiques où les longueurs des segments d'information croissent indéfiniment. La recherche des systèmes de codage permettant d'approcher au plus près la capacité maximale s'apparente à la quête du Saint Graal de la théorie de l'information [2]. Dans la pratique les paquets sont naturellement de taille bornée (souvent limitée à quelques milliers de bits) et les codes correcteur sont aussi de performances sous optimales.

Il existe deux sortes de corrections d'erreurs:

1. la correction d'erreurs sans intervention de l'émetteur;
2. la correction d'erreurs interactive avec l'émetteur.

Dans le second cas il y a une interaction entre le récepteur et l'émetteur pour corriger les erreurs détectées lors de la première transmission. En général cette correction passe par la simple re-transmission des segments erronés. La demande de retransmission passe par une voie de retour qui est en général un canal secondaire qui, lui aussi, est sujet à erreur. Dans ce cas on se place dans une architecture dite avec canal à *feedback* [4]. Dans le contexte du canal est à diffusion le premier canal et la voie de retour sont en général confondus mais ce n'est pas une règle absolue. En technologie des protocoles, on appelle *Automatic Repeat Query* (ARQ), les procédures de re-transmission.

Lorsque la correction d'erreurs a lieu entièrement sur le destinataire sans usage de voie de retour, on est en technique dite *Forward Error Correction* (FEC).

La figure 5.4 donne le format standard d'un paquet de données.

La zone FEC dans le paquet corresponde aux champs qui contiennent la redondance nécessaire pour la correction des erreurs de transmissions éventuelles. Un code correcteur d'erreurs corrige les erreurs en fonction de sa capacité. En général il ne peut pas corriger plus d'un nombre fixé k de blocs erronés dans un paquet. Notons que ces blocs erronés peuvent eux-mêmes se trouver dans les champs redondants FEC. Donc pour un taux d'erreur aléatoire ρ par bloc, la probabilité pour qu'un paquet ne soit pas corrigeable est:

$$p_{n,k}(\rho) = \sum_{i=k+1}^{n} \binom{n}{i} \rho^i (1-\rho)^{n-k} \tag{5.2}$$

où n est le nombre total de blocs dans le paquet. La quantité $p_{n,k}(\rho)$ est aussi la probabilité de rejet d'un paquet de données, dans le cas, fort fréquent, où le récepteur du paquet ne peut tirer d'utilité d'un paquet erroné, même partiellement. La quantité $\log \rho - \log p_{n,k}(\rho)$ exprimée en décibel (dB) s'appelle l'*efficacité* du code correcteur.

Au minimum le champ FEC est réduit à un CRC (*Checksum*) qui permet la détection des erreurs non corrigeables et permet au récepteur de décider du rejet du paquet erroné et du début d'une procédure ARQ.

On a donc une capacité pratique C_r du canal qui dépend des paramètres précédents et bien entendu $C_r < C_m$. Là-dessus se superpose une nouvelle inconnue qui est le *protocole de communication.*

La capacité du protocole de communication

Dans un réseau à accès multiples il importe que tous les utilisateurs connectés suivent une règle précise pour accéder à la ressource de communication. Une méthode idéale consisterait à disposer d'un gendarme électronique omniscient et doué d'ubiquité qui contrôlerait les accès en fonction exacte des besoins de chacun. Malheureusement ce type de contrôle n'existe pas

dans la pratique. Dans ce cas il faut prévoir un contrôle distribué, appelé *protocole* qui résulte en général en une nouvelle réduction de la capacité. La capacité C_p du canal qui reste après l'application du protocole, vérifie: $C_p < C_r$.

Il existe plusieurs types de protocoles d'accès comme le temps partagé périodique (TDMA pour *Time Division Multiple Access*) qui consiste à alouer périodiquement la même tranche de temps à chaque utilisateur connecté. Cette méthode est en général peu efficace et induit de lourds délais d'accès au canal. En effet les délais sont alors proportionnels au nombre total d'utilisateurs. Une alternative efficace réside dans les protocoles à accès aléatoire et à résolution de collisions. Quand deux utilisateurs transmettent en même temps, un seul paquet est susceptible de passer. Dans les réseaux câblés usuels les paquets émis simultanément se détruisent mutuellement: c'est une collision. Les protocoles à accès aléatoire présentent en général de bien meilleurs délais d'accès que le temps partagé périodique, sous réserve bien entendu d'une résolution efficace des collisions.

En fait la capacité C_p ne peut s'apprécier de manière valable que si l'on précise le modèle du trafic soumis au canal. Par exemple si le trafic est restreint sur un seul utilisateur on aura $C_p = C_r$ pour un protocole à accès aléatoire libre et $C_p = C_r/N$ pour le temps partagé périodique (N est le nombre total d'utilisateurs). Un modèle extrême consiste à envisager une population virtuellement infinie d'utilisateurs connectés générant un trafic total de λ paquets par unité de temps. Un protocole qui résiste au modèle de la population infinie est un protocole qui jouit de propriétés de stabilité et de performances très satisfaisantes.

Il existe des méthodes d'évaluation de la quantité C_p. La première est celle appliquée protocole Aloha, un des premiers réseaux de données à accès aléatoire. On suppose que le temps est partagé en tranches de tailles égales appelée slots. Sur chacun des slots les utilisateurs transmettent ou retransmettent indépendamment leur paquet avec chacun une probabilité p spécifiée à l'avance. Si le nombre d'utilisateurs susceptibles de transmettre est grand on estime que les transmissions effectives par slot suivent une loi de Poisson de paramètre μ (voir la section suivante pour la description de cette loi). Une transmission sans collision ayant lieu avec probabilité $\mu e^{-\mu}$ on a donc l'identité de débit

$$\mu e^{-\mu} = \lambda \tag{5.3}$$

Ce qui donne $C_p = \max_\mu\{\mu e^{-\mu}\} = e^{-1}$ paquets par slot. En fait cette évaluation est fausse parce que provenant d'une méthodologie trop grossière. Mais la quantité e^{-1} a longtemps été prise pour la borne supérieure de la capacité des protocoles à accès multiples sur canal sloté.

En passant on a donné la définition du protocole (ou canal) mono-slotté comme celui où le canal est slotté et où les paquets sont de tailles uniques égales à la durée du slot. Les utilisateurs ne sont autorisés à transmettre qu'au début des slots.

5.1.2 La modélisation de trafic et la loi de Poisson

En plus de la modélisation des sources d'information, on a vu que l'étude des canaux à diffusion et à accès multiples nécessite une nouvelle composante théorique qui est la modé-

lisation de trafic. Le trafic est le concept qui décrit l'ensemble des sources d'informations connectées au réseau et la répartition dans le temps de leur activité.

La loi de Poisson est aux modèles de trafic ce que la loi de Bernoulli est aux modèles des sources d'information: le modèle le plus simple dont la connaissance est un préalable indispensable.

La loi de Poisson est la manière la plus naturelle et la plus simple de décrire le processus de génération de paquets d'information issus de sources *nombreuses* et *indépendantes*. À l'origine, le mathématicien Poisson avait introduit sa loi pour expliquer la statistique des chutes de cheval dans la Grande Armée. Plus récemment la loi de Poisson a été utilisée pour modéliser les émissions de particules en radio-activité, les atomes remplaçant les chevaux et les neutrons, les cavaliers.

Revenons à la description de la loi de Poisson dans les réseaux de télécommunication. Supposons que le taux de génération de paquets d'une source soit de λ paquets par unité de temps, et considérons un intervalle de temps arbitraire de longueur Δt. La loi de Poisson établit que la probabilité pour que k paquets, k entier, soient générés durant l'intervalle de temps en question est

$$\frac{(\lambda \Delta t)^k}{k!} e^{-\lambda \Delta t} \ .$$

La loi établit aussi le fait que le processus de génération est sans mémoire, c'est à dire que ce qui se passe durant un intervalle de temps arbitraire est indépendant du passé (et de l'avenir). Une autre manière de décrire la loi de Poisson est de caractériser la durée aléatoire I séparant deux événements consécutifs. La probabilité, $\Pr\{I \geq t\}$, pour que cette durée soit plus grande qu'un réel positif t donné a pour expression

$$\Pr\{I \geq t\} = e^{-\lambda t} \ ,$$

ou, à l'échelle infinitésimale

$$\Pr\{I \in [t, t + \mathrm{d}t[\} = e^{-\lambda t} \lambda \mathrm{d}t \ .$$

Il y a plusieurs modèles de trafic basés sur la loi de Poisson. Il y a le modèle des n sources poissoniennes de taux respectifs λ_i paquets par slot (voir figure 5.5). Le taux cumulé est $\lambda = \sum_{i=1}^{i=n} \lambda_i$. On a vu aussi le très important modèle de la population infinie, à intensité cumulée λ (voir figure 5.6).

5.1.3 La borne de Kelly

Kelly [5], et moi-même indépendamment (non publié), avons calculé une borne supérieure de la capacité maximale pour un protocole mono-slotté à arrivées libres. Par protocole à arrivées libres on signifie que lorsqu'un utilisateur génère un nouveau paquet, il en tente la transmission sur le premier slot venu. Si cette transmission résulte en un échec (collision) alors l'utilisateur entre dans l'état de retransmission et applique un protocole de résolution de collisions.

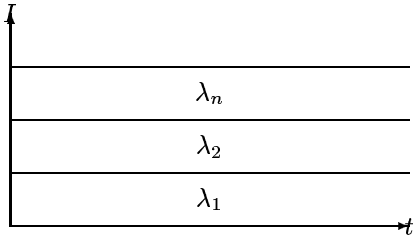

FIG. 5.5 – *Les n sources poissoniennes continues, temps t en abscisse, intensité de trafic I en ordonnée*

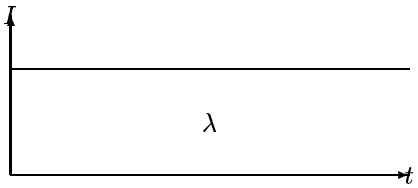

FIG. 5.6 – *Modèle de la population infinie, temps t en abscisse, intensité de trafic I en ordonnée*

On appelle α_0, α_1 et α_2 les proportions des slots qui voient respectivement aucun paquet en retransmission, un seul paquet en retransmission et deux paquets ou plus en retransmission. Les quantités α_0, α_1, α_2 dépendent évidemment du protocole de résolution de collisions.

On fixe naturellement

$$\alpha_0 + \alpha_1 + \alpha_2 = 1 \tag{5.4}$$

Comme la loi des arrivées libres est de Poisson de paramètre λ, la proportion des slots qui voient une transmission réussie, indépendamment de sa qualité d'arrivée libre ou de retransmission, est donc $\alpha_0 \lambda e^{-\lambda} + \alpha_1 e^{-\lambda}$. Si le protocole est stable et donne des délais d'accès finis on a égalité entre le taux d'arrivées et le taux de succès:

$$\alpha_0 \lambda e^{-\lambda} + \alpha_1 e^{-\lambda} = \lambda \tag{5.5}$$

On a évidemment $\lambda \leq 1$. De (5.5) on déduit:

$$(\alpha_0 + \alpha_1) e^{-\lambda} \geq \lambda \tag{5.6}$$

et par conséquent, avec (5.4):

$$e^{-\lambda} \geq \lambda \tag{5.7}$$

La simple condition numérique (5.7) implique $\lambda \leq \lambda_0$ avec $\lambda_0 e^{\lambda_0} = 1$. On a $\lambda_0 = 0.567143\ldots$.

Donc quelque soit le protocole mono-slotté à arrivées libres on a $C_p \leq \lambda_0 C_r$ dans le modèle de la population infinie.

M	C_p/C_r
2	0.360177
3	0.401599
4	0.399293
5	0.387241
6	0.373354
7	0.359731

FIG. 5.7 – *Performances des protocoles en arbres mono-slotté*

5.1.4 Les protocoles en arbre

On a déjà mentionné que la borne d'Aloha slotté, $e^{-1} = 0.367\ldots$, n'était pas valable. Une preuve en est apportée par les protocoles en arbre à arrivées libres dont on peut déterminer de manière exacte les capacités théoriques C_p. Les protocoles en arbre sont décrits en détail dans les articles consacrés aux modems câbles. Ils ont été introduits par Capetanakis [7] et Tsybakov-Mikhailov [8]. Ils donnent lieu à des évaluations exactes de performances par des développements analytiques [9]. Brièvement la résolution de collisions en arbre suit un arbre de résolution aléatoire dont les degrés de branchement sont tous égaux à un entier donné M. Le développement de l'arbre de résolution est très semblable à la structure d'un trie. Massey passe en revue de manière brillante les différentes versions du protocole en arbre dans [6]. La capacité du protocole en arbre à arrivées libres et population infinie dépend de la valeur de M et les valeurs sont données dans le tableau 5.7. On y constate que les protocoles en arbres dont le degré M est situé entre 3 et 6 présentent des capacités relatives C_p/C_r supérieures à e^{-1} [10].

5.1.5 Les protocoles à réservation

Les protocoles mono-slottés sont des cas extrêmes où les paquets sont tous de même taille minimale égale à la longueur du slot. Ce cas est typique avec les systèmes ATM où les données sont transmises dans des cellules de tailles toutes égales à 53 octets. Lorsque les paquets ont des tailles dépassant plusieurs slots on recourt avantageusement aux protocoles à *réservations*.

Dans le protocole à réservation, lorsque l'utilisateur a un paquet de données à transmettre, il crée d'abord un paquet de réservation. Le paquet de réservation fait un slot exactement et il contient comme information minimum la longueur du paquet de données auquel il est attaché. On appelle parfois le paquet de réservation, le jeton de réservation pour insister sur la petitesse relative dudit paquet.

Les jetons de réservation sont transmis comme dans un protocole mono-slotté et les collisions éventuelles donnent lieu à résolution de collision. Chaque fois qu'un jeton de réservation est transmis avec succès – donc reçu avec succès par les unités de contrôle du réseau, le canal de transmission est réservée de manière convenue pour l'utilisateur sur une période

M	C_p/C_r	$C_p(8)/C_r$	$C_p(16)/C_r$
2	0.360177	0.818296	0.900069
3	0.401599	0.842988	0.914806
4	0.399293	0.841713	0.914054
5	0.387241	0.834866	0.910002
6	0.373354	0.826580	0.905058
7	0.359731	0.818008	0.899894

FIG. 5.8 – *Capacités des protocoles en arbre avec réservation*

égale à la longueur du paquet attaché. Ladite longueur ayant été déclarée dans le jeton de réservation. L'utilisateur qui bénéficie de cette réservation aura l'exclusivité de cette période et transmettra sans collision les données de son paquet.

En conséquence le temps du canal est partagé alternativement entre

1. des périodes réservées pour les données;
2. des périodes à compétition pour les jetons de réservation.

Les performances du protocole à réservation dépendent du protocole de résolution de collisions pour les jetons : elles sont en général plus avantageuses que celles du protocole mono-slotté. Quand la taille moyenne L des paquets de données tend à augmenter, la capacité $C_p(L)$ pratique tend à se rapprocher de la capacité réelle C_r. On a la formule simple:

$$C_p(L) = \frac{L}{\frac{1}{C_p} + L} \tag{5.8}$$

où C_p est la capacité du protocole mono-slotté employé pour la transmission des jetons de réservation, appelé *protocole de réservation.*

Si le protocole de réservation est le protocole en arbre décrit précédemment et si on considère que les arrivées sont uniformes et poissoniennes durant les périodes à compétition on obtient les capacités pratiques listés dans le tableau 5.8

Le protocole à réservation avec résolution de collisions en arbre de degré $M = 3$ est la base du standard IEEE 802.14 pour les modems-câble. Le protocole est décrit et analysé en détail dans la section correspondante.

5.1.6 Les protocoles CSMA

Les protocoles CSMA (pour *Carrier Sense Multiple Access*) constituent une variante particulière des protocoles à réservation. Le début de chaque paquet joue le rôle du jeton de réservation. En effet à partir du moment où un utilisateur détecte le signal physique du début d'un paquet, celui-ci s'abtient de toute émission jusqu'à l'extinction du signal à la fin du paquet, protégeant ce dernier de toute collision. On suppose que le temps de détection du

signal d'un paquet prend une durée égale à un slot, correspondant au délai de la propagation plus intégration au niveau des récepteurs.

La différence entre les protocoles CSMA et les protocoles à réservations pures réside dans la possibilité résiduelle de collision sur la totalité du paquet de données. En effet dans le protocole CSMA deux utilisateurs peuvent encore transmettre leur paquet en commençant sur le même slot. Dans ce cas ils ne peuvent se détecter mutuellement et entrent en collision sur toute la longueur commune de leurs paquets, causant des dommages irréparables à leurs données qui nécessitent la retransmission. Dans le cas de la réservation pure seul le jeton de réservation aurait été perdu en gâchant moins de capacité du canal.

Si tous les paquets sont de même longueur L et que le trafic est uniforme de Poisson sur chaque slot (hors zone réservée) on obtient la capacité $C_p^{\mathrm{CSMA}}(L)$ du protocole CSMA:

$$C_p^{\mathrm{CSMA}}(L) = \frac{(\alpha_0 \lambda e^{-\lambda} + \alpha_1 e^{-\lambda})L}{\alpha_0 e^{-\lambda} + (1 - \alpha_0 e^{-\lambda})L} C_r \tag{5.9}$$

où α_0 et α_1 sont les paramètres de retransmission par slot déjà mentionnés dans l'établissement de la borne de Kelly. Lorsque L augmente indéfiniment on a la limite:

$$\lim_{L \to \infty} C_p^{\mathrm{CSMA}}(L) = \frac{\alpha_0 \lambda e^{-\lambda} + \alpha_1 e^{-\lambda}}{1 - \alpha_0 e^{-\lambda}} C_r \tag{5.10}$$

Si on reprend l'approximation grossière (mais Ô combien pratique) qui assume que le taux des paquets transmis par slot est de Poisson de paramètre μ on obtient

$$C_p^{\mathrm{CSMA}}(L) = \frac{\mu e^{-\mu} L}{e^{-\mu} + (1 - e^{-\mu})L} C_r \tag{5.11}$$

avec

$$\lim_{L \to \infty} C_p^{\mathrm{CSMA}}(L) = \frac{\mu e^{-\mu}}{1 - e^{-\mu}} C_r \tag{5.12}$$

On constate que la quantité ci-dessus tend vers C_r quand $\mu \to 0$.

Si les paquets ne sont pas tous de même longueur on a le facteur correctif:

$$C_p^{\mathrm{CSMA}}(L) = \frac{\mu e^{-\mu} L}{e^{-\mu} + \mu e^{-\mu} L + L^*(\mu)} C_r \tag{5.13}$$

avec

$$L^*(\mu) = \sum_k \frac{\mu^k}{k!} e^{-\mu} L_k$$

où L_k est l'espérance mathématique de la longueur du plus grand paquet parmi k paquets indépendants. En passant, on remarque que le dénominateur de $C_p^{\mathrm{CSMA}}(L)$ est la transformée de Poisson de la suite des L_k au point μ (avec la convention $L_0 = 1$).

5.1.7 Les protocoles CSMA/CD

Le protocole CSMA/CD (pour *CSMA with Collision Detection*) est une variante du protocole CSMA auquel on adjoint la possibilité de détecter les collisions dès le début de chaque paquet. Le protocole CSMA/CD est à la base des protocoles Ethernet (IEEE 802.3) et présente de bien meilleures performances que le protocole CSMA. Si une collision a lieu elle est détectée dès le début de chacun des paquets et les transmissions sont interrompues avant la fin du premier slot.

Le protocole CSMA/CD se comporte exactement comme un protocole à réservation pures – à la différence, mineure, que le début du paquet jouant le rôle de jeton de réservation, il contient déjà des données qu'il ne sera pas nécessaire de retransmettre si le paquet est transmis *in extenso*.

On a donc l'estimation:

$$C_p^{\mathrm{CD}}(L) = \frac{L}{\frac{1}{C_p} + L - 1} C_r \tag{5.14}$$

comme dans l'expression de la capacité du protocole à réservation. Le terme $L-1$ au dénominateur, au lieu de L provient du fait que le jeton de réservation contient déjà partiellement des données du paquet qu'il ne faut pas compter deux fois. Il est clair que $C_p^{\mathrm{CSMA}}(L) < C_p^{\mathrm{CD}}(L) < C_r$ et que $C_p^{\mathrm{CD}}(L)$ converge plus vite que $C_p^{\mathrm{CSMA}}(L)$ vers C_r quand L augmente.

5.2 Protocoles pour les modems câble

Il existe une demande pour des connexions Internet vraiment plus rapides qui se précise. En effet la généralisation des services multimédias en ligne ne peuvent plus se contenter des 28.8 kbps offerts par les modems téléphoniques, voire 64 kbps par les modem ISDN. Un accès ultra-rapide à Internet reviendrait, selon J. Waldhuter de Nynex Science et Technologie, à procurer un confort de connexion qui reviendrait "à pouvoir remplir sa baignoire à la demande en moins de cinq secondes".

Les technologies de l'accès rapide qui constitueraient actuellement des alternatives économiquement viables sont:

1. l'ADSL,
2. le satellite,
3. les modems câbles.

L'ADSL consiste à établir des connexions digitales rapides asymétriques sur des lignes téléphoniques normales. Elles nécessite l'installation de modems spécifiques chez l'abonné et dans les concentrateurs téléphoniques. Les débits dans le sens descendant sont estimés à 1.5-9 Mbps et dans le sens montant, 16-500 kbps, en fonction de la longueur et de la qualité de la ligne en cuivre existante. Il nécessite un réseau intermédiaire aussi à hauts débits entre les concentrateurs.

Le satellite offre une grande couverture à faible prix, il nécessite l'installation d'antennes collectives ou individuelles en plus du modem à la maison; ce qui revient à des investissements relativement faibles. Débit estimé descendant: 400 kbps ou plusieurs mégabit par seconde. Défaut principal: devoir recourir au téléphone ou au câble modem pour la voie de retour.

Le modem câble utilise le câble TV existant, et nécessite un modem chez l'abonné. Débits estimés descendants: 0.5-30 Mbps ; montants: 0.7-14 Mbps. Il est particulièrement bien adapté aux zones urbaines déjà couvertes par le câble.

Toutes les alternatives décrites ci-dessus demande la normalisation de protocoles d'accès performants. De nombreuses expérimentations ont été menées de part et d'autre de l'Atlantique.

Les principales caractéristiques physiques du réseau câblé sont:

a) l'existence d'un canal montant permettant l'interactivité;

b) un débit, 1-10 Mbps en montée et 10-30Mbps en descente, assurant la transmission symétrique à haut débit.

Le comité de normalisation international IEEE 802.14 s'occupe de la définition du protocole d'accès à la voie montante. Nous avons présenté des protocoles d'accès en arbre, dont l'un a été retenu pour les mécanismes de réservations à la voie montante. En particulier la norme s'appuie sur un mécanisme d'accès qui entrelace d'une manière naturelle les slots de requêtes avec les slots de données; il permet par conséquence un multiplexage optimal des sources avec des délais d'accès réduits. Le protocole de réservation est un algorithme en arbre légèrement modifié pour tenir compte de la détection différée des collisions et de l'entrelacement des réservations.

L'analyse du protocole a été obtenue sous divers modèles de trafic comme le trafic de Poisson continu ou le sursaut ponctuel de trafic [12, 13].

5.3 Les sources on/off

Jusqu'à présent le modèle de prédilection dans les évaluations d'algorithmes de télécommunications était le trafic uniforme de Poisson. De récentes statistiques de trafics, établies notamment à Bellcore (New Jersey) ont montré que certains trafics de données s'écartaient très sensiblement du modèle poissonnien. Des études récentes montrent que de tels trafics peuvent être modélisés par des superpositions de processus On/Off.

Les conséquences pratiques concernent essentiellement les capacités des mémoires-tampons au niveau des routeurs IP ou dans les commutateurs ATM. En effet des analyses de files d'attente simples montrent que les niveaux d'occupation des files d'attente suivent alors une loi à décroissance polynomiale alors que sous le modèle poissonien uniforme la décroissance est exponentielle.

L'analyse utilise les propriétés de la transformée de Mellin qui se révèle un outil efficace pour traquer les comportements asymptotiques polynomiaux. C'est donc une nouvelle application à cette transformée, qui depuis les brillants travaux de Philippe Flajolet est devenu un outil indispensable à l'analyse d'algorithmes.

On définit une source on/off par un processus de création de paquets qui passe alternativement de l'état actif (*on*) à l'état passif (*off*). À l'état passif, l'utilisateur ne génère aucun paquet; à l'état actif il génère des paquets comme une source poissonienne de λ paquets par unité de temps. Une source on/off modélise assez bien le comportement d'un utilisateur d'internet. Par exemple un *surfeur* du Net aura tendance à alterner de courtes périodes actives avec de courtes périodes passives, parce qu'il téléchargera de courts fichiers qu'il rejettera tout aussi rapidement tant qu'il n'aura pas fixé son attention sur une information significative. L'utilisateur avisé téléchargera tout de suite l'intégralité du fichier qu'il recherche pour l'étudier à son aise ensuite, avant de passer à une autre source d'information. Dans ce dernier cas on aura de longues périodes actives suivies d'encore plus longues périodes passives.

Une source on/off aléatoire a des périodes actives et périodes passives de durées aléatoires indépendantes. Une source on/off exponentielle présente des périodes actives (respectivement des périodes passives) distribuées de manière exponentielle de paramètre τ_1 (respectivement τ_0). En d'autre termes

$$\begin{aligned} \Pr\{\text{période active de durée supérieure à } t\} &= e^{-\tau_1 t} \\ \Pr\{\text{période passive de durée supérieure à } t\} &= e^{-\tau_0 t} \end{aligned}$$

Il apparaît que la source est sans mémoire en ce qui concerne la détermination des moments de transition entre l'état actif et l'état passif et *vice versa*. Cette propriété permet de manipuler de manière aisée la modélisation d'un tel trafic.

La durée moyenne de la période active est donc τ_1^{-1} et celle de la période passive est τ_0^{-1}. Donc la charge moyenne de la source on/off est $\frac{\tau_0}{\tau_0+\tau_1}\lambda$.

Dans ce qui suit on imagine un ensemble dénombrable de sources on/off exponentielles indépendantes dont le trafic cumulé doit transiter par un réseau de télécommunication (voir figure 5.9). Le résultat surprenant de mon étude [14] est que sous certaines conditions cet ensemble de sources on/off indépendantes et sans mémoire va générer un trafic cumulé sujet à des corrélations à long terme.

Soit $I(x)$ l'intensité de trafic à l'instant t, on a corrélation à long terme quand la covariance de $I(t)$ et $I(t+x)$ décroît asymptotiquement comme l'inverse une puissance de x, en x^β où β est identifié comme un coefficient d'*auto-similarité* de Hurst . On appelle $C(x)$ la covariance dans l'état stationnaire (dans ce cas le paramètre t s'efface). Pour une seule source on obtient:

$$C(x) = \lambda^2 \frac{\tau_0 \tau_1}{(\tau_0+\tau_1)^2} e^{-(\tau_0+\tau_1)x} \tag{5.15}$$

Pour un ensemble dénombrable de sources on obtient

$$C(x) = \sum_i C_i(x) \tag{5.16}$$

où $C_i(x)$ est la covariance exprimée pour la source i isolée. Comme les $C_i(x)$ sont des fonctions exponentielles, $C(x)$ est donc une somme harmonique d'exponentielles, analysable de manière avantageuse par transformée de Mellin. La transformée de Mellin de $C(x)$,

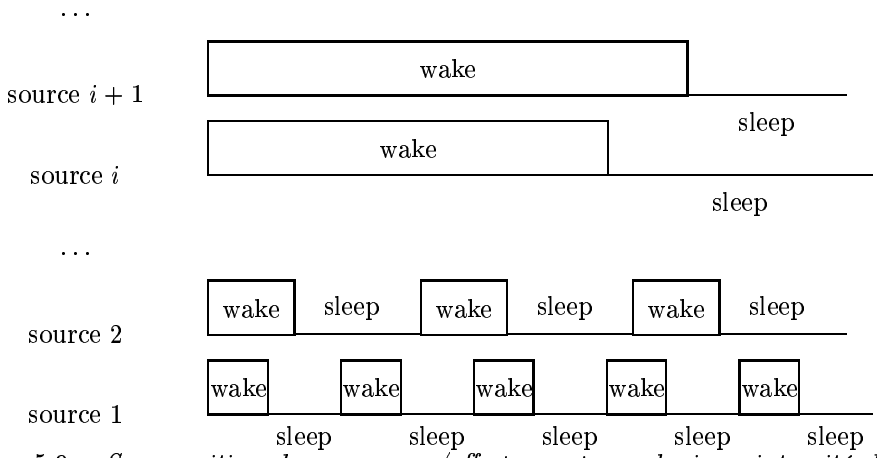

FIG. 5.9 – *Superposition de sources on/off, temps t en abscisse, intensité de trafic I en ordonnée*

$C^*(s) = \int_0^\infty x^{s-1} C(x) dx$ satisfait l'identité

$$C^*(s) = \Gamma(s)(\sum_i \lambda_i^2 \tau_{0,i} \tau_{1,i} (\tau_{0,i} + \tau_{1,i})^{2-s}) \ , \tag{5.17}$$

où on reconnaît dans le second membre une série de Dirichlet constituées des paramètres, λ_i, $(\tau_{0,i}, \tau_{1,i})$ des sources on/off. L'existence de poles dans les singularités de cette série de Dirichlet permet de déterminer le comportement polynomial asymptotique de $C(x)$ en utilisant la transformée de Mellin inverse et le théorème des résidus [11]. Par exemple si le premier pole de la série de Dirichlet est en $s = \beta$ avec résidu μ on obtient:

$$C(x) = \frac{1}{2i\pi} \int_{-i\infty}^{+i\infty} C^*(s) s^{-x} ds = \mu \Gamma(\beta) x^{-\beta} + O(x^{-\beta-\varepsilon} \ . \tag{5.18}$$

Un exemple de condition simple pour obtenir ces corrélations lourdes est que les deux suites des $\tau_{0,i}$ et des $\tau_{1,i}$ décroissent elles aussi comme l'inverse d'une puissance de i. Par exemple $(\tau_{0,i} + \tau_{1,i}) = i^{-\beta}$ pour $\beta > 1$ et $\tau_{0,i} \tau_{1,i} \lambda_i = i^{-\beta}$. Dans ce cas $C^*(s) = \Gamma(s)\zeta((1-s)\beta)$ (fonction *zeta* de Riemann) et a un pole simple en $s = 1 - 1/\beta$.

Une superposition de sources on/off aux caractéristiques divergentes est une manière efficace de modéliser la diversité des utilisateurs susceptibles d'être simultanément actifs sur un réseau de communication de grande dimension. Pour garder un système viable il faut

évidemment que la somme totale des charges moyennes $\frac{\tau_0}{\tau_0+\tau_1}\lambda$ reste inférieure à la capacité C_p du réseau. La mesure de l'impact sur les performances notamment au niveau des délais d'accès et des taux de pertes aux routeurs est une donnée cruciale pour la conception des protocoles de nouvelles générations.

5.4 Conclusion

Le domaine des communications est en explosion. Mais comme le problème à n corps, les modélisations sur des systèmes distribués sont assez ardues et il est parfois difficile de s'écarter des modèles de traffic simples comme le trafic uniforme de Poisson. C'est dans ce contexte que l'analyse algorithmique des protocoles permet de faire des percées interessantes.

Bibliographie

[1] C. SHANNON, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol 27, pp. 379-423, 623-656, Juillet-Octobre 1948.

[2] S. VERDU, "Fifty years of Shannon Theory," preprint, juin 1998.

[3] T. COVER, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol 20, pp. 2-14, 1972.

[4] N. GAARDER, J. WOLF, "The capacity region of a multiple-access channel with feedback," *IEEE Trans. Inform. Theory*, vol 21, pp. 100-102, 1975.

[5] F. KELLY, "Stochastic models of computer communication systems," *J. Royal Stat. Soc.*, B, 47, 1985.

[6] J. MASSEY, "Collision resolution algorithms and random-access communication," *Multi-User Communication Systems*, G. Longo, CISM courses and lectures 255, pp. 73-137, 1981.

[7] J. CAPETANAKIS, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inform. Theory*, vol 14, pp. 505-515, 1979.

[8] B. Tsybakov, V. Mikhailov, "Free synchronous packet access in a broadcast channel with feedback," *Probl. Inform. Transmission*, vol 14, pp. 259-280, 1979.

[9] G. FAYOLLE, P. FLAJOLET, M. HOFRI, P. JACQUET, "Analysis of a stack algorithm for random multiple access communication," *IEEE Trans. Inform. Theory*, vol 31, pp. 244-254, 1985.

[10] P. MATHYS, P. FLAJOLET, "Q-ary collision resolution algorithms in random access systems with free or blocked access communication," *IEEE Trans. Inform. Theory*, vol 31, pp. 217-243, 1985.

[11] P. FLAJOLET, X. GOURDON, P. DUMAS, "Mellin transforms and asymptotics: harmonic sums," *Theoretical Comp. Science*, vol 144, pp. 3-58.

[12] PHILIPPE JACQUET, PAUL MUHLETHALER, PHILIPPE ROBERT, "Asymptotic average access delay analysis: adaptive p-persistence versus tree algorithm," IEEE 802.14 working paper 96-248, 1996.

[13] P. JACQUET, P. MUHLETHALER, P. ROBERT, "Performant implementations of tree collision resolution on CATV network," IEEE 802.14 working paper 96-115, 1996.

[14] P. JACQUET, "Analytic Information Theory in Service of Queueing with Aggregated exponential On/Off Arrivals," 35th anniversary Allerton Conference, 1997.

Chapitre 6

Les réseaux sans fils

Libre comme l'air...

6.1 La problématique des réseaux de données mobiles

Imaginez des robots mobiles dans un site de production, dans l'espace, dans des lieux inaccessibles à l'homme ou trop dangereux. Ou bien les voitures "intelligentes" de demain ; ou bien les individus souhaitant "rester connectés" pendant leurs déplacements. Communiquer sans fil, voici le dénominateur commun à tous ces scénarios.

Mais contrairement à la radiotéléphonie et aux techniques cellulaires déjà connues, ces communications sont de type "multimedia" (voix, données, images) et doivent éventuellement être assurées sans infrastructure d'appoint (balises, stations centrales). De plus ces communications sont plus exigeantes en terme de performances. La brique essentielle d'un tel système de communication est le réseau numérique local par radio. Au premier abord un tel objet paraît simple à concevoir. Pourtant si cette brique est restée absente jusqu'à maintenant du domaine des réalisations concrêtes, c'est parce qu'elle pose des problèmes très ardus.

Le premier problème est d'ordre matériel. L'onde radio subit des dédoublements sur les obstacles qu'elle rencontre, ce qui en général altère fortement l'information qu'elle transporte quand le débit dépasse les millions de bits par seconde. Pour lutter contre ce phénomène on introduit des redondances massives dans le code transporté ce qui nécessite le recours à des composants très puissants.

Le deuxième problème est d'ordre algorithmique et concerne la communication des stations en réseau. Le problème est plus complexe que pour les réseaux locaux filaires. En effet le réseau est ouvert et mobile, le médium de communication est sujet à des perturbations constantes. Ces contraintes imposent le recours à des protocoles d'accès d'un type particulier qui feront appel à des algorithmes nouveaux tout en respectant les interfaces et les logiciels de communications existant.

Le troisième problème concerne la normalisation. Le spectre électromagnétique est une ressource très demandée. L'attribution des fréquences est un processus arbitré aux niveaux internationaux avec les implications politiques que l'on devine. Comme il est difficile d'imaginer plusieurs types de réseaux sans fil partageant le même milieu de communication, il est indispensable d'adopter l'approche la plus concertée possible.

Nous allons illustrer ces problèmes techniques dans les sections qui suivent.

6.2 La propagation et le traitement du signal

Les réseaux sans fils confèrent une liberté d'utilisation inégalée mais cette liberté ne doit pas masquer la quantité de problèmes non triviaux dont elle a nécessité la résolution.

Les réseaux sans fils restent des réseaux ouverts, dans la mesure où la propagation des ondes électromagnétiques est difficile à contraindre ou à limiter en portée. Il n'est pas envisageable de construire des cages de Faraday tout autour des portions de réseau que l'on désirerait isoler.

Le traitement du signal est la contrainte technologique majeure dans le déploiment des réseaux sans fil; l'excellent livre [3] analyse en détail les techniques de traitement.

6.2.1 La contrainte de la puissance reçue

L'atténuation des signaux avec la distance limite en premier la portée pratique des ondes radio. La puissance reçue $P(R)$ à une distance R décroît en raison inverse d'une puissance de R:

$$P(R) = \frac{A}{R^{\alpha}} \tag{6.1}$$

où α est le coefficient d'atténuation de la propagation. Dans le vide on a $\alpha = 2$. Dans l'air, dans un milieu matériel ou en présence d'une répartition aléatoire d'obstacles hétérogènes on a $\alpha > 2$. Si N est la valeur du bruit ambiant au récepteur on obtient une portée théorique R_m telle que

$$P(R_m) = K\mathrm{N} \tag{6.2}$$

et au delà de laquelle le signal reçu ne sera pas distinguable du bruit. La quantité K dépend de la technique de modulation et des paramètres du bruit. Il est à noter que le bruit ambiant est en général celui qui a lieu à l'intérieur du récepteur à cause de l'échauffement de ses composants. Le bruit ambiant suit un modèle statistique qui donne lieu à diverses variantes. Le modèle de bruit le plus connu et le plus simple est le modèle du *bruit blanc gaussien* où le bruit est modélisé selon des séquences i.i.d. de lois normales de moyenne nulle. La quantité N désigne alors la variance du bruit. Nous n'entrerons pas dans le détail de la modèlisation du bruit.

Shannon [1, 2] a montré que dans le cas du bruit blanc gaussien la capacité maximale entre un émetteur et un récepteur à distance R. ne peux pas dépasser la valeur C_m:

$$C_m = F\log(1 + \frac{P(R)}{\mathrm{N}}) \ , \tag{6.3}$$

où F est la largeur de bande affectée à la communication sans fil. Cette formule est en fait valable pour tous les types de communications soumis à des facteurs de puissance physiques.

Dans la pratique où les paquets sont limités en tailles et en codes de correction d'erreurs, la réception d'un paquet sans erreur n'a vraiment lieu que si le rapport *signal sur bruit* dépasse une certaine valeur critique. En général $K \approx 10$ est un bon ordre de grandeur. Donc

$$R_m = (\frac{A}{KN})^{1/\alpha} \ . \tag{6.4}$$

6.2.2 La contrainte des trajets multiples

Aux problèmes de puissance s'ajoutent les limitations dues aux trajets multiples et aux décalages entre les échos et qui introduisent au niveau du récepteur une sorte de bruit dépendant proportionnel à la puissance reçue $P(R)$ (voir figuretrajet. Contrairement à la problématique du signal sur bruit, on ne peut pas améliorer la situation en augmentant la puissance émise.

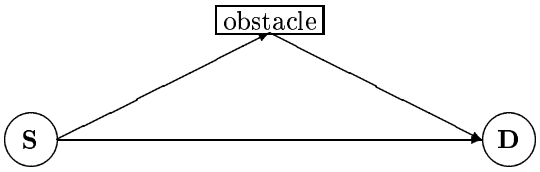

FIG. 6.1 – *Les trajets multiples*

On appelle *Décalage Inter-Symbole* (DIS) le ratio du décalage moyen sur la durée de l'émission d'un symbole [4]. *Grosso modo* la réception des données devient critique quand le DIS s'approche de l'unité, c'est à dire quand les symboles ont tendances à se mélanger avec ceux de leurs échos. Cette limitation, indépendante de la puissance émise, devient le facteur limitatif principal lorsqu'on passe aux réseaux sans fils à hauts débits (de l'ordre du Mbps).

Une modélisation raisonnable des trajets multiples en présence d'obstacles aléatoires consiste à exprimer le décalage moyen comme une puissance de la distance R entre l'émetteur et le récepteur. En général on prend la puissance un, qui est une sorte de pire cas. Tout autre puissance β comprise entre $1/2$ et 1 correspond à une sorte de disposition "fractale" des obstacles. Lorsque les obstacles sont disposées de manière homogène, la loi des Grands Nombres s'applique et $\beta \leq 1/2$.

Si pour un DIS maximal imposé on appelle C_r le débit maximal en symboles par seconde et R_i la portée maximale pour le débit en question on a

$$C_r(R_i)^\beta = \text{cte} \tag{6.5}$$

C_r	R_i
10 Mbps	10 - 100 m
1 Mbps	100 m - 1 km
100 kbps	1 - 10 km

FIG. 6.2 – *Portées pratiques en milieu urbain*

La constante dépend du DIS acceptable et donc de la complexité de la technique de décodage et du traitement du signal. Pour les techniques de décodages usuelles dans les milieux de type urbain on a les valeurs approximatives listées dans le tableau 6.2 pour le cas le pire $\beta = 1$. Il faut noter que la dernière ligne (100 kbps) correspond en gros à la capacité totale d'une fréquence du GSM (8 canaux temporels agrégés à 13 kbps chacun). Bien entendu ces valeurs subissent de fortes variations en relation avec la nature des environnements.

La complexité de la technique de décodage est fonction de la valeur du DIS maximal acceptable. Les techniques dites à *égalisation* (GSM, Hiperlan) nécessitent de *redresser* la réponse du canal de transmission par la résolution d'un système linéaire sur un échantillonnage des signaux mesurés lors de la réception du paquet. La dimension du système linéaire correspond au nombre d'échantillons nécessaire pour analyser la réponse du canal [5]. Ce nombre est en gros proportionnel à la valeur maximale acceptable du DIS. Donc la complexité du décodage est proportionnelle au carré du DIS acceptable. Si on appelle B la complexité de décodage par symbole on obtient donc lorsque B varie:

$$\frac{(C_r(R_i)^\beta)^2}{B} = \text{cte}_2 \tag{6.6}$$

Si on dispose d'une puissance de calcul maximale C_rB pour le traitement du signal, on obtient l'identité

$$(R_i)^{2\beta}(C_r)^3 = \text{cte}_3 \tag{6.7}$$

qui, pour ressembler à la loi de Kepler [7] sur les révolutions des planètes (mais avec l'échange des puissances), n'en prouve pas moins les énormes difficultés techniques qu'il faut résoudre lorsque l'on veut obtenir des hauts débits sur des portées significatives. En d'autres termes à puissance de traitement du signal égale, une multiplication par 10 du débit revient à diviser par 30 la portée réelle (pour $\beta = 1$).

En récapitulatif rapide, les contraintes des propagation des ondes reviennent à deux classes:

1. la contrainte du signal sur bruit;
2. la contrainte du décalage inter-symbole.

Bien que ces deux contraintes sont susceptibles d'avoir des effets croisés, par exemple un meilleur rapport signal sur bruit permet de traiter une plus grande valeur du DIS, il est raisonable en première approximation de traiter les effets de manière séparées. Dans ce cas

la portée efficace sera la valeur minimum des portées obtenues par la contrainte du signal sur bruit et par la contrainte du décalage inter-symbole.

6.3 Les performances d'un réseau sans fil

Cette section fait écho à un exercice d'expertise auprès de l'ETSI (*European Telecommunication Standard Institute*) au sujet des performances des protocoles de communication pour le trafic aérien.

Il n'est pas aisé de définir les performances d'un réseau sans fil. La difficulté majeure vient de la nature ouverte, non bornée de ces réseaux. Par exemple deux émetteurs éloignés ont la possibilité d'émettre simultanément sur la même fréquence sans entrer en collision. C'est ce qu'on appelle la *réutilisation spatiale*. En contrepartie un récepteur ne sera pas en mesure de recevoir un paquet émis de trop loin et il faudra sans doute procéder à des répétitions de proche en proche pour relayer le paquet vers sa destination lorsque celle-ci n'est pas à portée directe.

Les éléments précédents constituent les points de différence essentiels avec les réseaux câblés dans lesquels la connectivité est assurée à toute distance du fait de l'atténuation négligeable des signaux dans un milieu conducteur linéaire.

6.3.1 Paramètres d'intéret et modélisation

Comme le réseau est non borné on fait l'hypothèse d'une densité de trafic par unité de volume et unité de temps. En fait on regarde plutôt une densité de trafic par unité de surface. En ce qui concerne le trafic aérien on regarde la densité de trafic estimé à l'intérieur de cylindres verticaux dans l'air, de rayon variable et de hauteur de 30 000 pieds. On appellera λ la densité de trafic en paquet par unité de surface et de temps. On suppose que le trafic et les sources de trafic sont répartis dans le temps et l'espace suivant une loi uniforme de Poisson.

Ayant défini la densité de trafic, il reste à définir la portée pratique moyenne R_p et l'aire moyenne σ_p où le paquet est reçu correctement. On appelle capacité unitaire le produit $\lambda\sigma_p L$, dénoté C_u et où L est la longueur moyenne des paquets. On restreint l'analyse à un canal unique, donc un récepteur ne peut recevoir qu'un seul paquet à la fois. On a par conséquent l'inégalité:

$$C_u \leq C_r \tag{6.8}$$

Les phénomènes de collisions et de bruit sont les facteurs qui vont limiter le coefficient d'utilisation en dessous de sa valeur maximale théorique.

Pour déterminer le coefficient d'utilisation il est avantageux d'introduire la fonction de portée $p(R)$ qui exprime la probabilité pour qu'un récepteur à distance R de l'émetteur reçoive correctement son paquet. On a les identités simples:

$$R_p = 2\pi \int_0^\infty p(r)\mathrm{d}r \tag{6.9}$$

et

$$\sigma_p = 2\pi \int_0^\infty rp(r)\mathrm{d}r \ . \tag{6.10}$$

On fait l'hypothèse simplificatrice du protocole mono-sloté telle que décrite dans le chapitre précédent. Le slot correspond à l'unité de temps. Pour simplifier davantage l'analyse on ignore les protocoles de retransmission en cas de collision, les paquets détruits ne sont pas retransmis.

Par contre on prend en compte les phénomènes de capture. Pour qu'un paquet soit reçu avec succès par un récepteur donné il faut que la puissance avec laquelle il est reçue soit K fois supérieure à la somme des puissances reçues simultanément des autres émetteurs. Nous ignorons la contribution due au bruit ambiant.

6.3.2 Analyse sous le modèle avec atténuation α

Dans ce modèle nous prenons un coefficient d'atténuation α strictement plus grand que deux, qui correspondrait à la propagation dans le vide. On définit la variable aléatoire W qui est la somme des puissances reçues sur l'antenne d'un récepteur aléatoire lors d'un slot arbitraire. Nous introduisons la transformée de *Laplace-Stieljes* $w(t)$ de la distribution de W, comme une fonction de variable complexe:

$$w(t) = \mathrm{E}[e^{-tW}] \tag{6.11}$$

En utilisant l'indépendance des contributions provenant des émetteurs situés sur des cercles concentriques différents autour du récepteur, tire l'expression intégrale:

$$w(t) = \exp(2\pi\lambda \int_0^\infty (\exp(-tx^{-\alpha}) - 1)x\mathrm{d}x) \ . \tag{6.12}$$

Le calcul donne la formule close

$$w(t) = \exp(-\pi\Gamma(1 - \frac{2}{\alpha})t^{2/\alpha}) \tag{6.13}$$

où $\Gamma(.)$ est la fonction *Gamma* d'Euler. Soit $F(x)$ la fonction de répartition de W: $F(x) = \Pr\{W > x\}$. En tenant compte des phénomènes de capture on a l'identité formelle

$$p(r) = 1 - F(\frac{1}{Kr^\alpha}) \ . \tag{6.14}$$

L'expression (6.13) peut être utilisée pour déterminer le comportement asymptotique de $F(x)$ quand x tend vers l'infini:

$$F(x) = \lambda\pi x^{-2/\alpha} - \lambda^2 \sin(\frac{4\pi}{\alpha})\Gamma(\frac{4}{\alpha})(\Gamma(1 - \frac{2}{\alpha}))^2 x^{-4/\alpha} + O(x^{-6/\alpha}) \tag{6.15}$$

Cette évaluation s'effectue en deux temps. Dans un premier temps on utilise la transformée de Laplace inverse:

$$F(x) = \frac{1}{2\pi} \int_{-i\infty}^{i\infty} w(t) \frac{e^{tx}}{t} \mathrm{d}t \tag{6.16}$$

et ensuite la transformée de Mellin $F^*(s)$ de $F(x)$:

$$\begin{aligned} F^*(s) &= \int_0^\infty F(x) x^{s-1} \mathrm{d}x \\ &= \frac{\sin(\pi s)\Gamma(s)}{\pi} \int_0^\infty w(t) t^{-s-1} \mathrm{d}t \\ &= \frac{\sin(\pi s)}{\pi s} \Gamma(s)\Gamma(1 - \frac{2}{\alpha}s) \left(\pi\lambda\Gamma(1 - \frac{2}{\alpha}) \right)^{2s/\alpha} \end{aligned}$$

L'analyse des singularités de la transformée de Mellin permet de déterminer de manière classique le développement asymptotique de la fonction d'origine $F(x)$. En revenant sur l'expression (6.14) nous en déduisons le développement asymptotique quand r tend vers zéro:

$$p(r) = 1 - \lambda\pi K^{2/\alpha} r^2 + O(r^3) \tag{6.17}$$

Cette expression peut servir à déterminer la portée pratique sous la contrainte d'un taux maximal de perte imposé. L'aire moyenne de réception a une expression exacte qui est

$$\sigma_p = \frac{2\sin(2\pi/\alpha)\lambda}{\alpha K^{2/\alpha}} \ . \tag{6.18}$$

Le débit unitaire a bien sûr pour expression:

$$C_u = \frac{2\sin(2\pi/\alpha)}{\alpha K^{2/\alpha}} C_r \tag{6.19}$$

Il est à noter que l'aire de réception tend vers zéro lorsque α tend vers deux, c'est à dire lorsqu'on se rapproche des conditions de propagations dans le vide. C'est un effet typique d'*horizon* où l'atténuation du signal ne contrecarre pas suffisamment l'augmentation des émetteurs distants dont la contribution cumulée devient infinie et empêche toute capture du paquet émis même par un émetteur proche. Nous allons traiter dans la prochaine sous-section le cas de la propagation dans le vide mais en imposant un horizon borné.

6.3.3 Analyse sous le modèle de la propagation dans le vide avec horizon

Lorsque $\alpha = 2$ la plupart des intégrales calculées précédemment divergent. Donc il est nécessaire d'introduire un horizon R_h au delà duquel un récepteur ne peut pas recevoir de signal d'un émetteur. Pour le trafic aérien la rotondité de la terre joue le rôle d'horizon:

en deçà d'une altitude de 30 000 pieds les avions sont masqués si la distance qui les sépare dépasse 400 nautiques.

La transformée de Laplace-Stieljes de la variable W a maintenant pour expression:

$$w(t) = \exp(2\pi\lambda \int_0^{R_h} (\exp(-tx^{-2}) - 1)x\mathrm{d}x) \tag{6.20}$$

qui peut être réécrite en:

$$w(t) = \exp(-\pi\lambda t\psi(\frac{t}{R_h^2})) \tag{6.21}$$

avec $\psi(z) = \int_z^\infty (1 - e^{-x})\frac{\mathrm{d}x}{x^2}$.

Le cas pratique, notamment pour le trafic aérien, intervient quand la valeur de l'horizon R_h est grande. Dans ce cas, plus précisément quand la *charge sous horizon* $\lambda\pi R_h^2$ est importante, on obtient l'estimation asymptotique:

$$F(x) = \frac{\lambda\pi}{x - \lambda\pi\log(\lambda\pi R_h^2)} + O(\frac{1}{x^2\log^2(\lambda\pi R_h^2)}) \tag{6.22}$$

dont la précision augmente quand x augmente en proportion avec le logarithme de la charge sous horizon. De cette estimation il découle quand r tend vers zéro:

$$p(r) \approx 1 - \frac{\lambda\pi K r^2}{1 - \lambda\pi K r^2\log(\lambda\pi R_h^2)} \tag{6.23}$$

L'aire moyenne de réception a l'expression asymptotique

$$\sigma_p \approx \frac{1}{\pi K\lambda\log(\lambda\pi R_h^2)} \tag{6.24}$$

La portée moyenne et le débit unitaire peuvent être dérivés de la même manière.

6.3.4 Exercice commenté: période des hellos

Le suivi des routes des avions est la préoccupation majeure du contrôle aérien. Pour ce faire les avions signalent périodiquement leur présence en diffusant un paquet *hello*. Le paquet hello contient les informations sur la route, la position et l'altitude de l'appareil. Recevant chacun des hellos des autres avions les pilotes peuvent gérer individuellement leur espace aérien en pointant les positions de chacun des avions voisins.

La période T d'émission des paquets hellos est imposée selon des règles de sécurité draconiennes. Compte tenu d'un débit unitaire donnée C_u, la taille s des paquets hello, quel est le nombre moyen d'avions pointés par pilote par période? Appelons N_p ce nombre.

Tout d'abord il n'est pas surprenant que le nombre moyen d'avions pointés ne dépendent pas de la densité μ des avions dans le ciel. En fait l'aire moyenne de réception est inversement proportionnelle à la charge du trafic, donc inversement proportionnelle à la densité des émetteurs périodiques.

Pour des raisons de symétrie évidentes le nombre moyen d'avions pointés est égal au nombre moyen d'avions dans l'aire de réception d'un paquet hello arbitraire. Comme $\lambda = \frac{\mu}{T}$ et $N_p = \sigma_p \mu$, on obtient

$$N_p = \frac{C_u}{C_p} T \tag{6.25}$$

si T est exprimée en multiple du slot. La quantité s étant la taille d'un slot, on peut écrire

$$N_p = \frac{C_u T}{s} \ . \tag{6.26}$$

En conclusion il y a un compromis entre la détermination de la période de rafraichissement des positions et le nombre d'avions pointables. L'analyse prouvait que le nombre d'avions pointés tournait autour de 7 à l'approche des aéroports. Ce nombre insuffisant plaide pour le recours à des fréquences multiples pour ce type de contrôle aérien.

6.4 La détection et la résolution de collision dans HIPERLAN

L'INRIA a contribué avec succès à la définition de la partie protocole de la norme HIPERLAN. Notre équipe, principalement constituée de Philippe Jacquet, Pascale Minet, Paul Mühlethaler, Nicolas Rivierre a été représentée à toutes les réunions de normalisation HIPERLAN organisées par l'ETSI.

Nous avons présenté et fait adopter le protocole d'accès à signalement actif qui marque l'originalité du standard HIPERLAN. Le signalement actif consiste à faire précéder l'émission du paquet par une série de périodes d'émission et de réception aléatoirement disposées. Les périodes d'émission ne contiennent aucune donnée, leur rôle consistant simplement à éliminer de la compétition les utilisateurs qui en détectent l'énergie. Les utilisateurs survivants gagnent le droit de transmettre leur paquet juste après leurs périodes de signalement actif. Les utilisateurs éliminés s'abstiennent d'émettre jusqu'à la fin des paquets transmis par les survivants.

Si il y a plus d'un seul survivant, les paquets ont de fortes chances d'entrer en collision, les récepteurs ayant peu de chance de se trouver en configuration de capture favorable. L'enjeu du signalement actif consiste à rendre cette éventualité la plus rare possible.

La technique d'évitement de collision par signalement actif est une amélioration notable de la technique par signalement passif, parfois connue sous la dénomination CSMA/CA (pour *Collision Avoidance*). Avec le CSMA/CA l'émission du paquet est précédée par une unique période de réception de durée aléatoire. Les vainqueurs de la compétition sont ceux qui sélectionnent la plus courte période de réception préalable et qui donc éliminent les autres compétiteurs dés que ceux-ci détectent l'énergie de leur transmission.

La technique d'accès par signalement actif peut être considérée comme une sorte de synthèse de travaux sur les protocoles en arbre adaptés aux spécificités de la transmission radio [8]. Ce travail a été particulièrement riche puisqu'il a permis un heureux mélange

d'une activité scientifique de haut niveau et d'un transfert de technologie réussi. Par exemple l'étude du comportement du protocole d'accès s'apparente au problème combinatoire classique de la sélection d'un vainqueur dans un tournoi aléatoire. L'analyse mathématique précise fournit des éléments qualitatifs et quantitatifs qui sont aussi clairement mis en évidence dans des simulations poussées, et qui permettent de qualifier le protocole.

6.5 Le protocole de routage dans HIPERLAN

Il existe trois types d'architectures sans fil:

- l'architecture *ad-hoc*: les utilisateurs communiquent entre eux en vue directe;
- l'architecture avec *station de base*: les utilisateurs communiquent en faisant transiter leurs paquets par une station de base en vue directe de tous les utilisateurs;
- l'architecture avec *routage interne*: les utilisateurs communiquent en faisant transiter leurs paquets sur une série d'utilisateurs intermédiaires.

Nous définissons la *fiabilité* d'un réseau comme la probabilité d'avoir la totalité des couples d'utilisateurs capables de communiquer entre eux. Des raisonnements simples permettent de calculer la fiabilité des trois types d'architectures sans fil. Supposons qu'un réseau sans fil de n utilisateurs repose sur le graphe complet à n sommets. Supposons que chaque arête ait la même probabilité p, indépendante, d'être défectueuse.

La fiabilité du réseau *ad-hoc* est donc égale à la probabilité d'avoir au moins une arête défectueuse, c'est à dire $(1-p)^{(n-1)n/2}$. La fiabilité du réseau avec station de base est $(1-p)^n$: il suffit que les n arêtes à la station de base fonctionnent correctement. La fiabilité du réseau à routage interne est plus difficile à calculer. Elle correspond en fait à la probabilité pour que le graphe construit avec arêtes valides ne contienne qu'une seule composante connexe. Étant donné une probabilité p donnée, les études sur la connectivité des graphes [6] permettent de calculer la fiabilité du réseau avec routage.

Le tableau ci-dessous donne des exemples de fiabilités dans le cas où $p = 0.1$:

Type	8	16	32
ad-hoc	0.53144	0.05233	0.00001
base	0.65610	0.43047	0.18532
routage	0.99581	0.99999	0.99999

Nous avons aussi défini le protocole de routage [9] interne qui permet au réseau HIPERLAN de fonctionner indépendamment de toute infrastructure câblée, sans restriction de topologie et avec une fiabilité accrue. Nous nous sommes inspirés des principes établis et accumulés sur le routage depuis les premiers travaux sur Arpanet et nous avons introduit des idées nouvelles permettant d'optimiser et de tenir compte de la spécificité d'un réseau sans fil.

Le routage de paquets dans des réseaux sans fil est rendu particulièrement délicat du fait de la nature versatile des transmissions radio. Les algorithmes de routage utilisés dans

HIPERLAN prennent par exemple en compte la possibilité de liens asymétriques. D'autres particularités intéressantes de l'algorithme de routage d'HIPERLAN sont relatives à l'optimisation du routage des paquets à destinations multiples (*multicast*). Par ailleurs l'implémentation du protocole de routage permet une représentation réduite de la topologie qui permet un calcul des tables de routage en $0(N)$, où N désigne le nombre de nœuds du réseau. Elle permet aussi des optimisations par rapport à l'occurrence des évènements de type rupture de liens.

6.6 Conclusion

Le domaine des réseaux sans fils ajoute à la théorie des réseau et des systèmes distribués une composante suplémentaire qui est la modélisation non triviale de la propagation des ondes. Cela en fait un domaine neuf où il est illusoire de croire que l'on fera des percées significatives en appliquant juste à la lettre les recettes des réseaux filaires. Par exemple la problématique du routage sans fil s'écarte sensiblement de celle de la théorie des graphes dans le routage filaire. En effet les transmissions radio sur des liens différents sont susceptibles d'interférer fortement. Dans ce domaine on peut dire qu'il y a encore beaucoup à faire et on peut se réjouir du moindre petit pas en avant.

Bibliographie

[1] C. SHANNON, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol 27, pp. 379-423, 623-656, Juillet-Octobre 1948.

[2] C. SHANNON, "Communication in presence of noise," *Proc. IRE*, vol 37, pp. 10-21, 1949.

[3] E. LEE, D. MESSERSCHMITT, *Digital Communication*, seconde édition, Kluwer, 1993.

[4] L. DOSSI, G. TARTARA, "Statistical analysis of measured impulse response functions of 2.0 GHz indoor radio channels," *IEEE Selec. Areas in Comm.*, vol 14, pp. 405-410, 1996.

[5] L. VANDENDORPE, "MIMO DFE equalisation for multitone DS/SS systems over multipath channels," *IEEE Selec. Areas in Comm.*, vol 14, pp. 405-410, 1996.

[6] I. GOULDEN, D. JACKSON, *Combinatorial Enumeration*, Wiley, 1993.

[7] J. KEPLER, *De Harmonia Mundi*, 1619.

[8] P. JACQUET, P. MINET, P. MÜHLETHALER, N. RIVIERRE, "Priority and Collision Detection with active Signaling: The Channel Access Mechanism of HIPERLAN," in *Wireless Personal Communications* Vol 4, No 1, pp. 11-25, 1997.

[9] P. JACQUET, P. MINET, P. MÜHLETHALER, N. RIVIERRE, "Increasing reliability in cable-free Radio LANs: Low level forwarding in HIPERLAN," in *Wireless Personal Communications* Vol 4, No 1, pp. 51-63, 1997.

Conclusion

It is not over until it is over...

Ce travail tente de dégager une perspective générale dans mes travaux de recherche. J'ai délibérément placé cette perspective dans celle de la théorie de l'information pour les raisons que j'ai expliquées en introduction. J'espère seulement au terme de ce mémoire que je n'aurais pas trop lassé mes lecteurs. Je pense de manière tout à fait égoïste que sa rédaction aura été une épreuve bénéfique.

J'ai essayé de montrer que l'utilisation des méthodes analytiques permet d'atteindre des évaluation précises et efficaces. Dans le tableau suivant je présente un récapitulatif rapide des étapes présentées dans ce mémoire.

rubrique	modèles	analyse complexe
Entropie	sources	poissonisation séries génératrices
Compression et structures de données	sources	séries génératrices équa. diff. analyse asympt.
Protocoles Communication	trafic	probabilités Mellin
Réseaux sans fil	propagation géométrie	Laplace séries génératrices

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399